

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): MIKI, et al.
Serial No.: Not yet assigned
Filed: January 5, 2004
Title: STORAGE DEVICE SYSTEM HAVING BI-DIRECTIONAL
COPY CONTROL FUNCTION
Group: Not yet assigned

LETTER CLAIMING RIGHT OF PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

January 5, 2004

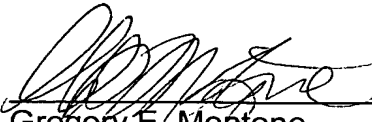
Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s) hereby claim(s) the right of priority based on Japanese Patent Application No.(s) 2003-128163, filed May 6, 2003.

A certified copy of said Japanese Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP



Gregory E. Montone
Registration No. 28,141

GEM/alb
Attachment
(703) 312-6600

日 本 国 特 許 庁
JAPAN PATENT OFFICE

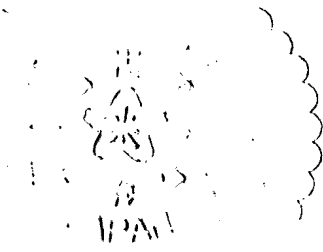
別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 3 年 5 月 6 日
Date of Application:

出 願 番 号 特 願 2 0 0 3 - 1 2 8 1 6 3
Application Number:
[ST. 10/C] : [J P 2 0 0 3 - 1 2 8 1 6 3]

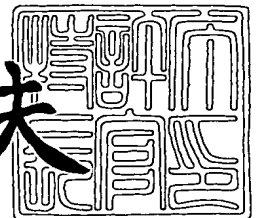
出 願 人 株式会社日立製作所
Applicant(s):



2 0 0 3 年 9 月 3 0 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫



出証番号 出証特 2 0 0 3 - 3 0 8 0 0 9 2

【書類名】 特許願

【整理番号】 NT03P0175

【提出日】 平成15年 5月 6日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 12/00

【発明者】

【住所又は居所】 神奈川県小田原市中里 3 2 2 番地 2 号 株式会社日立製作所 S A N ソリューション事業部内

【氏名】 三木 健一

【発明者】

【住所又は居所】 神奈川県小田原市中里 3 2 2 番地 2 号 株式会社日立製作所 S A N ソリューション事業部内

【氏名】 阿知和 恭介

【特許出願人】

【識別番号】 000005108

【氏名又は名称】 株式会社日立製作所

【代理人】

【識別番号】 100068504

【弁理士】

【氏名又は名称】 小川 勝男

【電話番号】 03-3661-0071

【選任した代理人】

【識別番号】 100086656

【弁理士】

【氏名又は名称】 田中 恭助

【電話番号】 03-3661-0071

【選任した代理人】

【識別番号】 100094352

【弁理士】

【氏名又は名称】 佐々木 孝

【電話番号】 03-3661-0071

【手数料の表示】

【予納台帳番号】 081423

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 双方向コピー制御機能を有する記憶装置システム

【特許請求の範囲】

【請求項 1】

複数の上位装置と、少なくとも 1 つの前記上位装置から書き込みデータを受け取る複数の記憶装置システムとを有するコンピュータシステムにおける前記記憶装置システムであって、

前記記憶装置システムは、複数の前記記憶装置システムに亘って同一のデータが保存されるように制御される論理ボリュームに対応して前記論理ボリューム上のデータを記録する物理記憶装置と、前記上位装置から前記書き込みデータを受け取った受領時刻を保存する手段と、前記論理ボリュームへの書き込みデータと対応する前記受領時刻とを他の前記記憶装置システムへ送信し他の前記記憶装置システムから書き込みデータと対応する前記受領時刻とを受信する通信インタフェース手段と、前記上位装置から受領した書き込みデータと前記通信インタフェース手段を介して受信した書き込みデータとが前記物理記憶装置の同じ格納場所に重複して書き込まれるときに前記受領時刻の順に書き込まれるように、前記論理ボリュームへの書き込みデータを対応する前記受領時刻から所定時間以上一時記憶装置上に待機させた後に前記物理記憶装置に書き込むよう制御するデータ一貫性保持制御手段とを有することを特徴とする記憶装置システム。

【請求項 2】

前記記憶装置システムは、さらに前記一時記憶装置に待機中の各書き込みデータに対応する前記受領時刻を前記受領時刻の古いものから順に配列したテーブルと、前記テーブルの先頭から順に、受領時刻から前記所定時間以上経過した書き込みデータを探索する手段とを有することを特徴とする請求項 1 記載の記憶装置システム。

【請求項 3】

前記記憶装置システムは、さらに書き込みデータの各ブロックが前記一時記憶装置に存在するか否かをビット値で設定するビットマップテーブルと、前記ビットマップテーブルを参照して新しい書き込みデータが他の書き込みデータと前記

同じ格納場所に重複して書き込まれるか否かを判定する手段とを有することを特徴とする請求項 1 記載の記憶装置システム。

【請求項 4】

前記記憶装置システムは、さらに前記上位装置から前記論理ボリュームの一部領域をロックする要求を受領して前記一部領域をロックする手段と、前記通信インタフェース手段を介して受領した前記ロック要求を他の前記記憶装置システムへ送信する手段と、前記通信インタフェース手段を介して他の前記記憶装置システムからロック要求を受領して指定された一部領域をロックする手段と、前記一部領域に対する前記上位装置及び他の前記記憶装置システムからの書き込みデータの要求を、当該一部領域をロックした上位装置からの要求である場合を除いて拒絶する手段とを有することを特徴とする請求項 1 記載の記憶装置システム。

【請求項 5】

複数の上位装置と、少なくとも 1 つの前記上位装置から書き込みデータを受け取る複数の記憶装置システムとを有するコンピュータシステムにおける前記記憶装置システムであって、

前記記憶装置システムは、複数の前記記憶装置システムに亘って同一のデータが保存されるように制御される論理ボリュームに対応して前記論理ボリューム上のデータを記録する物理記憶装置と、前記上位装置から前記書き込みデータを受け取った受領時刻を保存する手段と、前記論理ボリュームへの書き込みデータと対応する前記受領時刻とを他の前記記憶装置システムへ送信し他の前記記憶装置システムから書き込みデータと対応する前記受領時刻とを受信する通信インタフェース手段と、前記上位装置からの前記書き込みデータ及び他の前記記憶装置システムからの前記書き込みデータの各々に対応する前記受領時刻の古いものから順に配列したテーブルと、前記テーブルを参照し、前記論理ボリュームへの書き込みデータについて前記受領時刻から所定時間以上経過した書き込みデータを前記受領時刻の古い順に前記物理記憶装置に書き込むよう制御するデータ一貫性保持制御手段とを有することを特徴とする記憶装置システム。

【請求項 6】

前記記憶装置システムは、さらに前記上位装置から前記論理ボリュームの一部

領域をロックする要求を受領して前記一部領域をロックする手段と、前記通信インタフェース手段を介して受領した前記ロック要求を他の前記記憶装置システムへ送信する手段と、前記通信インタフェース手段を介して他の前記記憶装置システムからロック要求を受領して指定された一部領域をロックする手段と、前記一部領域に対する前記上位装置及び他の前記記憶装置システムからの書き込みデータの要求を、当該一部領域をロックした上位装置からの要求である場合を除いて拒絶する手段とを有することを特徴とする請求項5記載の記憶装置システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、記憶装置システムに係わり、特にある記憶装置システムから他の記憶装置システムへのコピーを双方向に行う技術に関する。

【0002】

【従来の技術】

上位装置であるホストコンピュータ及び複数の記憶装置システム（ストレージシステム）を有する情報システムにおいて、記憶装置システム間でのデータコピーを行う技術として、リモートコピーがある。リモートコピーとは、情報処理システムにおいて、上位装置が介在することなしに、物理的に離れた場所にある複数の記憶装置システムの間でデータのコピー、すなわち2重書きを行う技術である。記憶装置システムとは、複数の記憶装置及びそれらの記憶装置を制御するコントローラとから構成されるシステムを指す。リモートコピーを行う情報処理システムは、物理的に離れた場所にそれぞれ配置される記憶装置システムが専用回線あるいは公衆回線でお互いに接続されている。この接続に用いられる専用線あるいは公衆回線をリモートコピーラインという。

【0003】

ある記憶装置システムが有する論理的な記憶領域（以下、「論理ボリューム」という）のうち、リモートコピーの対象となる論理ボリューム（以下、「コピー元論理ボリューム」という）の容量と同容量の論理ボリュームがコピーされる他の記憶装置システムに確保される。この確保された論理ボリューム（以下、「コ

ピー先論理ボリューム」という)は、コピー元論理ボリュームと一対一の対応関係(以下、「コピーペア」という)を有するように形成される。コピー元論理ボリュームのデータは、専用回線または公衆回線を介してコピー先論理ボリュームにコピーされる。コピー元論理ボリュームに含まれるデータの更新があった場合、更新されたデータは専用回線等を介してコピー先論理ボリュームを有する記憶装置システムに転送され、コピー元論理ボリュームに対応するコピー先論理ボリュームにも更新されたデータが書き込まれる。リモートコピーの技術を用いると、複数の記憶装置システムを有する情報処理システムにおいて、複数の記憶装置システムで同一内容の論理ボリュームを保持する事ができる。

【0004】

コピー元とコピー先というコピーペアを形成することによって、コピー元からコピー先へとコピー方向が一方方向に決定される。コピー元論理ボリュームには、上位装置より書き込み処理ができる。逆にコピー先論理ボリュームは、上位装置の書き込み処理はできない。

【0005】

仮にコピーの方向を一方方向に固定せずにコピーペアを構成している記憶装置システムへの書き込みを可能とすると、それぞれの記憶装置システム内のデータの内容を同一に保つことができない。これは上位装置が記憶装置に書き込みを行い、その後に書き込み内容をコピー先の記憶装置に書き込むときのコピーデータの転送に要する転送時間のためである。

【0006】

具体的に2つの記憶装置システムA、記憶装置システムBの間でコピーペアが形成されている環境を例にとって説明を行う。ここで記憶装置システムAと記憶装置システムBは地理的に十分離れており、上位装置より記憶装置システムAにデータが書き込みされ、次に記憶装置システムAから記憶装置システムBへ2重化データを転送し書き込むまでに例えば1秒以上時間がかかるものとする。

【0007】

ここではほぼ同時刻に記憶装置システムAと記憶装置システムBの同じ領域にそれぞれの上位装置から異なる内容(それぞれ内容A、内容B)の書き込みがあっ

たとすると、記憶装置システムA、記憶装置システムBには、それぞれ内容A、内容Bが書き込まれる。それぞれ書き込み終了後、ほぼ同じタイミングで記憶装置システムAから記憶装置システムBへ、また記憶装置システムBから記憶装置システムAへ、内容A、内容Bが転送される。このような状況では、記憶装置システムA、記憶装置システムBが転送データを受領後、記憶装置システムA、記憶装置システムBにそれぞれ内容B、内容Aが書き込まれることになる。記憶装置システムA内は内容Aの上に内容Bが上書きされた状態となり、記憶装置システムB内は内容Bの上に内容Aが上書きされた状態となる。このような場合、記憶装置システムA、記憶装置システムB内の書き込まれた内容は異なったものとなり、ボリュームの2重化が正常に行われていないことになってしまう。

【0008】

このような状態を避け、ボリュームの完全な二重化を実現するために、コピー先、コピー元というようにコピー方向が一方方向に決められる。リモートコピーに関する技術は、米国特許5,742,792号公報（特許文献1）に開示されている。

【0009】

従来、複数の上位装置から共有される記憶装置は、任意の上位装置からの共有排他制御要求に基づき、個々の上位装置からのアクセス要求に対する共有排他制御を実現している。例えば上位装置と記憶装置とのインターフェースとしてSCSI (Small Computer System Interface)を採用する情報システムは、SCSIで規定するリザーブ系のコマンドを用いて、前記記憶装置の論理ボリューム単位での共有排他制御を実現することができる。ある上位装置が論理ボリュームをリザーブした際には、リザーブした上位装置からのみのリードアクセス、ライトアクセスが可能な状態となる。

【0010】

SCSIリザーブ系コマンドにおいて、ディスクのブロック単位で共有排他制御を行えるような拡張コマンドも用意されている。この論理ボリューム上の一部の領域（エクステント）をリザーブするSCSIコマンドはエクステント・リザーブ（以下、「領域リザーブ」という）と定義されている。リザーブされる領域

は、リザーブ属性をもつ。リザーブ属性は、リード・シェア、排他ライト、排他リード、排他アクセスが可能である。SCSI-2に関する技術は、1997年2月1日第3版CQ出版SCSI-2詳細解説の項目6.15（非特許文献1）で説明されている。

【0011】

現状のリモートコピー技術環境では、SCSIのリザーブ系コマンドによる共有排他制御機構は考慮されておらず、ある記憶装置システム中の論理ボリュームをリザーブコマンドによりロックした際も他の記憶装置システム内のリモートコピー対応論理ボリュームにはロックの状態が伝わらない。

【0012】

【特許文献1】

米国特許5,742,792号公報

【非特許文献1】

SCSI-2詳細解説、第3版、項目6.15、CQ出版、1997年2月1日発行

【0013】

【発明が解決しようとする課題】

上記従来技術のリモートコピーは、コピー元論理ボリュームのみにしか上位装置より書き込みができないという問題があった。またリモートコピー対応論理ボリュームにはリザーブ系コマンドによるロック状態が伝わらないという問題があった。

【0014】

本発明の第1の目的は、コピーペアを構成する記憶装置システム間でコピー方向を一方向に固定せず、双方向にコピーを行うよう制御することにある。

【0015】

本発明の第2の目的は、双方向コピーの下でリザーブ系コマンドによるリザーブ状態をリモートコピーが実施されている記憶装置システム間で伝播させることにある。

【0016】

【課題を解決するための手段】

本発明は、記憶装置システム間での双方向コピーを可能とするために、コピーペアを構成する記憶装置システムにデータ一貫性保持制御手段を設ける。このデータ一貫性保持制御手段は、上位装置から受領した書き込みデータと通信インタフェース手段を介して他の記憶装置システムから受信した書き込みデータとが物理記憶装置の同じ格納場所に重複して書き込まれるときに上位装置から書き込みデータを受領した受領時刻の順に書き込まれるように、コピーペアを形成する論理ボリュームへの書き込みデータを対応する受領時刻から所定時間以上一時記憶装置上に待機させた後に物理記憶装置に書き込むよう制御する。

【0 0 1 7】

また本発明の記憶装置システムは、さらに上位装置から論理ボリュームの一部領域をロックする要求を受領してその一部領域をロックする手段と、通信インタフェース手段を介して受領したロック要求を他の記憶装置システムへ送信する手段と、通信インタフェース手段を介して他の記憶装置システムからロック要求を受領して指定された一部領域をロックする手段と、この一部領域に対する上位装置及び他の記憶装置システムからの書き込みデータの要求を、当該一部領域をロックした上位装置からの要求である場合を除いて拒絶する手段とを有する。

【0 0 1 8】**【発明の実施の形態】****(1) 第 1 の実施例**

以下、双方向コピーに係る第 1 の実施例について図面を参照して説明する。

【0 0 1 9】

図 1 は、本実施形態のコンピューターシステム 1100 の構成図である。コンピューターシステム 1100 は、記憶装置システム 1070 に S A N (Storage Area Network) 1040 を介して接続された複数の上位装置 1000、上位装置 1010 からなるサイト 1110 と、記憶装置システム 1080 に S A N 1050 を介して接続された複数の上位装置 1020、上位装置 1030 からなるサイト 1120 で構成されている。記憶装置システム 1070 と記憶装置システム 1080 は、専用回線あるいは公衆回線を用いたリモートコピーライン 1060 で接続されている。記憶装置システム 1070 及び記憶装置システム 1080

は、リモートコピーライン1060を介してS C S I プロトコルを用いて互いに通信することができる。

【 0 0 2 0 】

図 2 は、図 1 で示すコンピューターシステム1100において双方向コピーが実施されている状態を説明する図である。図 2 において、上位装置1010は、記憶装置システム1070に対してデータの書き込み B 1200を行う。書き込み B 1200のデータは記憶装置システム1070に記憶された後、リモートコピーライン1060を介して、データが送信され（矢印1240）、記憶装置システム1080にコピーされる。同様に、上位装置1020から記憶装置システム1080への書き込み C 1210も、リモートコピーライン1060を介したデータ送信（矢印1230）が行われ、記憶装置システム1070にコピーされる。同様に、上位装置1030から記憶装置システム1080への書き込み D 1220も、リモートコピーライン1060を介したデータ送信（矢印1250）が行われ、記憶装置システム1070にコピーされる。つまり各上位装置1010、1020及び1030のそれぞれの書き込み B 1200、C 1210及びD 1220はそれぞれのサイトの記憶装置システムに書き込まれ、次に上位装置を介さずに相手サイトの記憶装置システムにコピーされる。

【 0 0 2 1 】

図 3 は、記憶装置システム1070のハードウェア構成を示している。なお記憶装置システム1070は、ディスクアレイ装置や、半導体記憶装置などを含んでいてもよい。記憶装置システム1070は、ホスト I / F 1300、D K C I / F 1320、ディスク制御部1350、共有メモリ1360、キャッシュメモリ1340、これらの間を通信可能に接続するクロスバススイッチなどで構成されるスイッチング制御部1330、タイマー1310、物理ディスク1370及びプロセッサ1380などで構成される。

【 0 0 2 2 】

ホスト I / F 1300は、C P U、メモリを備え、少なくとも 1 つの上位装置との間の通信を制御する。ホスト I / F 1300は、上位装置からのデータ I / O 要求を受信してそのデータ I / O 要求を共有メモリ1360に書き込む。なおリモートコピーの機能は、D K C I / F 1320の C P U がこの機能を実現するプログラムを実行することにより提供される。

【 0 0 2 3 】

キャッシュメモリ1340は、主としてホスト I / F 1300、D K C I / F 1320とディスク制御部1350との間で授受されるデータを一時的に記憶するために用いられる。例えばホスト I / F 1300が上位装置から受信したデータ入出力コマンドが書き込みコマンドである場合には、ホスト I / F 1300は上位装置から受信した書き込みデータをキャッシュメモリ1340に書き込む。またディスク制御部1350はキャッシュメモリ1340から書き込みデータを読み出して物理ディスク1370に書き込む。

【 0 0 2 4 】

ディスク制御部1350は、C P U、メモリを備え、ホスト I / F 1300やD K C I / F 1320より共有メモリ1360に書き込まれた I / O要求を読み出してその I / O要求に設定されているコマンド（本実施形態ではS C S I規格のコマンド）にしたがって、物理ディスク1370にデータの書き込みや読み出しなどの処理を実行する。ディスク制御部1350は、読み出しコマンドの場合は、物理ディスク1370から読み出したデータをキャッシュメモリ1340に書き込む。またデータの書き込み完了通知や読み出し完了通知などをホスト I / F 1300に送信する。ディスク制御部1350は、物理ディスク1370をいわゆるR A I D（Redundant Array of Inexpensive Disks）方式のR A I Dレベル（例えば、0、1、5）に従って、複数の物理ディスクに一つの論理ボリュームを分散配置する機能を備えることもある。

【 0 0 2 5 】

物理ディスク1370は、例えばハードディスク装置など書き込みデータを最終的に記録する物理記憶装置である。物理ディスク1370は記憶装置システムと一体型とすることも出来るし別筐体とすることもできる。D K C I / F 1320は、C P U、メモリを備え、他の記憶装置システムとの間でデータ転送をするための通信インターフェースであり、リモートコピーにおける他の記憶装置システムへのデータの転送は、このD K C I / F 1320を介して行われる。それぞれの記憶装置システムは1つのタイマー1310を持ち、それぞれのタイマーは、できる限り同一の時刻になるように調整されているものとする。タイマー1310はホスト I / F 1300が上位装置より I / Oの受付がなされた時刻を記録するためなどに用いられる。

【 0 0 2 6 】

プロセッサ1380は、CPU、メモリを備え、ホスト I / F 1300、DKC I / F 1320、ディスク制御部1350以外の後述するプログラムを実行する。

【 0 0 2 7 】

図 4 は、本実施形態の双方向リモートコピーに関するソフトウェア構成を示す図である。本構成は、記憶装置システムの双方向リモートコピーを実現するために、メイン制御2020、キャッシュ部2050およびデータ一貫性保持制御部2040の各プログラムを備える。これらのプログラムは、プロセッサ1380によって実行される。またメモリ上にビットマップテーブル2030を備える。

【 0 0 2 8 】

メイン制御2020は、ホスト I / F 1300から入出力要求を受け取り、データ一貫性保持制御部2040及びキャッシュ部2050を起動し、入出力処理の結果をホスト I / F 1300へ返す。またDKC I / F 1320を介する入出力要求の受け渡しを制御する。

【 0 0 2 9 】

ビットマップテーブル2030は、リモートコピーの対象である物理ディスク上の1ブロックが1ビットに対応するようビットマップ化したテーブルである。ビットマップテーブル2030は、ビット値 0、ビット値 1 の 2 つの状態を持つ。値 0 は、該当ビットに対応するディスクのブロックのデータがキャッシュメモリ1340上にキャッシュされていない状態を表す。値 1 は、該当ビットに対応するディスクのブロックのデータがキャッシュメモリ1340上にデータがキャッシュされており、キャッシュ上に最新のデータがある状態を表す。

【 0 0 3 0 】

キャッシュ部2050は、ホスト I / F 1300及びDKC I / F 1320からの書き込みデータをキャッシュメモリ1340に書き込む処理を行うプログラムである。キャッシュには書き込みデータのキャッシュ以外に、リードキャッシュがある。リードキャッシュ技術は、上位装置が記憶装置システム内のデータをリードする際に、物理ディスク1370へ直接アクセスし参照データをリードするよりも素早くリードデータを上位装置へ受け渡しができるように、キャッシュメモリ1340上に、頻繁

にアクセスのあるデータをキャッシュデータとして持つ技術である。しかし本実施形態では、本発明の特徴を明確にするためにキャッシュ部2050は、書き込みデータのキャッシュ制御のみ行うものとする。

【0031】

データ一貫性保持制御部2040は、入出力要求がデータ書き込みに関するものであるときに起動され、ビットマップテーブル2030に基づいて双方向コピーが実施される記憶装置システムの間でデータの一貫性が保持されるように制御する。

【0032】

ホスト I / F 1300 、 D K C I / F 1320からの記憶装置システムへの書き込みデータは、キャッシュメモリ1340上で一定時間保持された後に、ディスク制御部1350を介して物理ディスク1370に書き込まれる。データ一貫性保持制御部2040は、書き込みデータをキャッシュに保持する時間を監視制御する。この時間は、書き込みデータがコピー先に送信される転送時間及び本発明における制御処理にかかる時間を考慮した十分長い時間とする。

【0033】

本実施例では、ホスト I / F 1300が上位装置からの書き込み要求を受けた時刻より、3分間はキャッシュメモリ1340上に保持する。データ一貫性保持制御部2040は、3分以上キャッシュメモリ1340上にある書き込みデータを、コピーペアを構成する記憶装置システム間ではほぼ同じ時刻に一斉に物理ディスク1370へ書き込むよう制御する。この一斉書き込みは、1分ごとに行われる。例えば時刻が00時00分00秒、00時01分00秒、00時02分00秒のとき一斉書き込みが行われる。つまり3分以上4分未満キャッシュメモリ1340上でキャッシュされたデータは、時刻00秒のディスクへの書き込み処理の際に、物理ディスク1370へ書き込まれる。以下、この1分ごとに行われる物理ディスク1370への書き込み処理を「一斉書き込み処理」と呼ぶ。

【0034】

物理ディスク1370に書き込み内容を書き込む際には、コピーペアの記憶装置システム内のタイマー1310は、できる限り同一の時刻になるよう時刻合わせがされており、物理ディスク1370に書き込む内容は3分後にはまったく同じものとなる

ように制御される。これによって各記憶装置システムは、同一データをほぼ同じタイミングで物理ディスク1370に書き込む。

【0 0 3 5】

図5は、データー貫性保持制御部2040が保持するデーター貫性保持テーブル100のデータ形式を示す図である。データー貫性保持テーブル100は、テーブル管理番号101、受領時刻102、上位装置認識番号103、対象ブロック開始アドレス104、対象サイズ105、ストレージシリアル番号106及びキャッシュデータ格納アドレス107という項目によって構成される。

【0 0 3 6】

受領時刻102は、ホスト I / F 1300が上位装置より書き込みデータを受領した時刻を格納する。上位装置認識番号103は、記憶装置に書き込みを行った上位装置の識別番号である。この識別番号は上位装置ごとに一意であり、IPアドレス、ファイバーチャネルについて用いられるWWN (World Wide Name) などである。対象ブロック開始アドレス104及び対象サイズ105は、それぞれ書き込みの対象のブロック番号及び書き込みブロック数である。ストレージシリアル番号106は、記憶装置システムごとに付けられている一意の値であり、どの記憶装置システムが上位装置より書き込みの要求を受領したかを表す値となる。キャッシュデータ格納アドレス107は、書き込みデータが格納されているキャッシュメモリ1340のアドレスである。キャッシュデータ格納アドレス107は、C言語のmallocなどを用いてキャッシュメモリ1340上にデータ格納領域が確保されるとき、その先頭アドレスを示す（キャッシュ上のデータの削除は、C言語のfreeなどによって行える）。

【0 0 3 7】

データー貫性保持テーブル100は、受領時刻102を基準に時系列順にソーティングされており、最新のレコード（あるいはエントリ）がテーブルの一番末尾に来るようになっている。テーブル管理番号101は、テーブルの先頭より1から順に2、3、…と整数の管理番号を格納している。管理番号1は、キャッシュメモリ1340上で一番時刻の古いデータを表すレコードであり、管理番号が一番大きなレコードは最新の時刻をもつ書き込みレコードとなるように、レコードが配列され

る。管理番号が一番大きなレコードの次のレコードは、未登録のエントリなのでその管理番号に－1を代入する。

【 0 0 3 8 】

具体的に図5を用いて説明すると、現在100個のレコードがデーター貫性保持テーブル100に登録されている。各レコードは受領時刻102により時間の古いもの昇順でソートされている。テーブル管理番号101は、先頭レコードの番号を1として順に100まで割り当てられている。テーブル管理番号101が100の一つ下のレコードは存在しない。この未登録のレコード格納領域のテーブル管理番号101には－1が代入されている。

【 0 0 3 9 】

図6は、ビットマップテーブル2030のデータ形式を示す図である。ビットマップテーブル2030の各欄は、物理ディスク上のブロックに対応し、ブロックの順に配列されている。各欄にはビット値1又は0が設定され、上記のようにそのブロックのデータがキャッシュされているか否かを示す。一時ビットマップテーブル200は、上位装置からの入出力要求が入出力の対象とするブロックの範囲のビットマップを格納するテーブルである。

【 0 0 4 0 】

図7及び図8を用いてメイン制御部2020の処理手順について説明する。メイン制御部2020は、ホストI/F1300及びDKCI/F1320から受け取った入出力要求（以下I/Oという）を処理する。ホストI/F1300からのI/Oは、上位装置からのI/Oであり、DKCI/F1320からのI/Oは、他の記憶装置システムからのI/Oである。ここでホストI/F1300からの要求に関する処理は図7で示し、DKCI/F1320からの要求に関する処理は図8で示す。

【 0 0 4 1 】

図7は、ホストI/F1300からのI/Oを処理するメイン制御部2020の処理手順を示すフローチャートである。ホストI/F1300から記憶装置システムにI/O要求が来たとき、ステップ3000で参照系のコマンド（SCSIにおけるリードなどのコマンド）か、更新変更系のコマンド（SCSIにおけるライトなどのコマンド）かを識別し、処理を分岐させる。本実施例においては、参照系のコマン

ドと更新変更系のコマンドのみに注目する。参照系コマンドのときは、ステップ3005に処理が移る。更新変更系のコマンドのときは、ステップ3050に処理が移る。

【0042】

ステップ3005では、上位装置が要求している参照範囲（参照のブロック開始アドレス、ブロック数）に対応するビットマップテーブル2030のビット値を参照し、それらを一時ビットマップテーブル200にコピーする。一時ビットマップテーブル200作成後、処理はステップ3010に移る。ステップ3010では、一時ビットマップテーブル200の示すブロックアドレスにおいて、最新の内容がキャッシュメモリ1340上に存在するか、物理ディスク1370上に存在するかを判断する。

【0043】

上位装置が要求している参照範囲において、一時ビットマップテーブル200のビット値0のブロック範囲と、ビット値1のブロック範囲では、最新のデータの格納場所が変わってくる。ビット値が0のブロック範囲の参照要求は、物理ディスク1370上のデータが最新のデータという状態であり処理はステップ3020に移る。ビット値が1のブロック範囲の参照要求は、上位装置の参照範囲はキャッシュメモリ1340上のデータが最新のデータという状態であり処理はステップ3040に移る。ステップ3020は、ディスク制御部1350を介して参照範囲データを読み込む処理である。ステップ3040は、キャッシュ部2050を介して参照範囲データを読み込む。キャッシュ部2050を介してデータを参照する際は、データ一貫性保持テーブル100を用いて参照対象範囲のデータを読み込む。具体的には、データ一貫性保持テーブル100のレコードをテーブル管理番号101の一番大きなものから小さいものへ順番に、参照範囲と対象ブロックアドレス204、対称サイズ205を比較していき、変更のあったデータがキャッシュされているキャッシュメモリ1340のアドレスを検索し、キャッシュメモリ1340より参照範囲データを読み込む。

【0044】

ステップ3020、ステップ3040の後、ステップ3030でそれぞれの読み込みデータを一時ビットマップテーブル200のブロック範囲の読み込みデータとして結合させ、ホスト I/F 1300に渡す。ホスト I/F 1300にデータを渡した後に I/O 処

理が終了する。ホスト I / F 1300は、渡された読み込みデータを参照要求が発行された上位装置へ送信する。

【 0 0 4 5 】

ステップ3050、ステップ3060及びステップ3070は、ホスト I / F 1300から更新変更コマンドを受理した際の処理である。記憶装置システム内のデータに書き込みを行う際には、データー貫性制御部2040が処理を行うので、メイン制御部2020は、データー貫性保持制御2040にデータを渡し、リモートコピー先記憶装置システムへのデータ送信、ホストへの更新変更処理の完了を通知する処理となる。リモートコピー先記憶装置システムへのデータ送信の場合、メイン制御部2020は、ライトされた内容に加えて、データー貫性保持テーブル200の項目である受領時刻102、上位装置識別番号103、対象ブロック開始アドレス104、対象サイズ105、ストレージシリアル番号206の情報を送信する。ここで送信方法について特に記載しないが、S C S I のライトコマンドを拡張したベンダ固有コマンドを用いてもよい。

【 0 0 4 6 】

ステップ3050は、データー貫性保持制御部2040の処理であり、詳細は後述する。ステップ3050後にステップ3060に処理が移り、書き込みデータがD K C I / F 1320へ渡される。D K C I / F 1320は、この書き込みデータをコピー先記憶装置システムのD K C I / F 1320へ送信する。その後ステップ3070に処理が移行する。ステップ3070は、ホスト I / F 1300へ書き込み完了報告を行う。ホスト I / F 1300は、書き込み要求を行った上位装置に書き込み完了の通知を行う。

【 0 0 4 7 】

図 8 は、メイン制御部2020がD K C I / F 1320から書き込みデータを受け取ったときの処理のフローチャートである。ステップ3100は、データー貫性保持制御部処理である。

【 0 0 4 8 】

図 9 は、データー貫性保持制御部2040が上位装置からの書き込み要求を処理する際のフローチャートである。データー貫性保持制御部2040は、リモートコピーペアを形成する論理ボリュームが、完全に2重化されたボリュームとなるように

ライトデータ（書き込み内容）の処理を行う制御部である。データー貫性保持制御部2040は、データー貫性保持テーブル100とビットマップテーブル2030を更新し、ライトデータをキャッシュメモリ1340上へ書き込む処理を行う。

【0049】

ステップ3200は、ライトデータの書き込み範囲（ディスクのブロック開始アドレス、ブロック数）に対応するビットマップテーブル2030の範囲にビット値1のビットがあるか否かを調べる。この結果、ライトデータの範囲のビットマップがすべてビット値0の際は、処理はステップ3270に移る。ライトデータの範囲に一部分でもビット値1がある場合（キャッシュ上にライトデータの書き込みブロックと同じブロックのデータがキャッシュされている場合）、処理はステップ3210に処理は移る。

【0050】

ステップ3210では、データー貫性保持制御部2040は、ライトされたブロック数分のビットを持つ一時ビットマップテーブル200を作成する。一時ビットマップ200は、ライトデータの書き込みブロックと対応しており、一時ビットマップ200の先頭はライトデータの書き込み先頭ブロックと対応している。ステップ3210では、ビットマップテーブル2030を参照し用意した一時ビットマップテーブル200にライトデータの範囲のビット値をコピーする。コピー後、処理はステップ3220に移る。

【0051】

ステップ3220は、データー貫性保持テーブル100の一番最近に書かれたレコードを検索する。一番最近書かれたレコードの検索は、データー貫性保持テーブル100を上から順に検索してゆき、テーブル管理番号が-1となったエントリの一つ上のエントリとなる。検索されたレコードのテーブル管理番号101をメモリ上のカレントという変数に入れる。カレントは、整数を代入できる変数とする。テーブル管理番号をカレントに代入後、処理はステップ3230に移る。

【0052】

ステップ3230からステップ3260の処理はループ処理となっており、ループ終了条件は一時ビットマップテーブル200のビット値がすべて0になることである。

以下、ビット値を 0 にしていく手順を説明する。一時ビットマップテーブル 200 のビット値が 1 に対応するブロックへの書き込みデータは、既にキャッシュメモリ 1340 上に存在している。よってデータ一貫性保持テーブル 100 のレコードを検索することにより、このキャッシュメモリ 1340 上の重複データ（一時ビットマップテーブル 200 においてビット値 1 であるデータ）をすべて検索し、検索で見つかったレコードの重複範囲に対応する一時ビットマップテーブル 200 のビット値を 0 にする。また重複データに対応する範囲に後述する重複ブロック削除処理 3240 を行う。このデータ一貫性保持テーブル 100 の各レコードを検索していくことによって、一時ビットマップテーブル 200 のビット値を 0 としていく処理が行われる。一時ビットマップテーブル 200 のビット値がすべて 0 になるとループ処理は終了する。

【 0 0 5 3 】

ステップ 3230 は、データ一貫性保持テーブル 100 のテーブル管理番号 201 が変数カレントと一致するレコード（以下「カレントレコード」という）の対象ブロック開始アドレス 104 と対象サイズ 105 を用いて、ライトデータの書き込み範囲と重複する部分があるか否か判定する。判定の結果、重複部分がない場合、処理はステップ 3250 に移る。重複がある場合、処理はステップ 3240 に移る。ステップ 3240 の重複ブロック削除処理終了後はステップ 3250 に移る。ステップ 3250 は、比較対象レコードであるカレントレコードを 1 つ上のレコードに変更する処理である。具体的にカレントの変数値を - 1 してやればよい。変数の値を変更後、処理はステップ 3260 へ移る。ステップ 3260 は、ループ処理の終了条件判定を行う処理であり、一時ビットマップテーブル 200 の値がすべて 0 になっているかを判定する。

【 0 0 5 4 】

一時ビットマップテーブル 200 のビット値がすべて 0 になっている場合には、処理はステップ 3270 へ移る。ビット値に 1 が残っている場合には、ループの先頭であるステップ 3230 に処理が移る。ステップ 3270 は、ライトデータの書き込み範囲に相当するビットマップテーブル 2030 のビット値をすべて 1 に変更する処理である。処理終了の後、ステップ 3280 に移る。ステップ 3280 では、データ一貫性保持制御部 2040 は、ライトデータに関する情報をデータ一貫性保持テーブル 100 に

書き込む。具体的には、テーブル管理番号101に一番最近書かれたレコードのテーブル管理番号+1の値を書き込む。その一行下のテーブル管理番号101には-1を書き込む、更にライトデータの情報を元に、受領時刻102、上位装置識別番号103、対象ブロック開始アドレス104、対象サイズ105、ストレージシリアル番号106及びキャッシュデータ格納アドレス107をそれぞれ書き込む。データ一貫性保持テーブル100を更新した後、処理はステップ3290に移る。ステップ3290は、キャッシュメモリ1340上のキャッシュデータ格納アドレス107に設定したアドレスにライトデータを書き込む。書き込み終了後、データ一貫性保持制御部2040の処理が終了する。

【0055】

図10は、重複ブロック削除処理を示すフローチャートである。重複ブロック削除処理は、記憶装置システムに書き込み要求のあるライトデータの書き込み範囲と、キャッシュメモリ1340上にあるキャッシュデータの書き込み範囲に重複がある場合の処理である。重複があるということは、ライトデータがキャッシュデータを上書きするか又はキャッシュデータがライトデータを上書きする。重複ブロック削除処理は、この重複した範囲において上書きされ更新される部分を削除する処理である。

【0056】

(a) ライトデータがキャッシュデータを上書きする場合

リモートコピーを用いない環境など、通常はホスト I / F 1300で受領したライトデータの順番にキャッシュ書き込みが行われる。よってライトデータが最新のデータとなり、ライトデータがキャッシュデータを上書きする。

【0057】

(b) キャッシュデータがライトデータを上書きする場合

D K C I / F 1320で受領したライトデータは、最新の書き込みデータである保証がない。これはリモートコピー元記憶装置システムにライトデータが書き込まれた後に、リモートコピーライン1060を介して転送されてきたライトデータであり、すでに受領時刻102より時間が経っている。よってキャッシュデータの方がライトデータよりも最新のデータとなるケースが生じる。このケースの場合、キ

キャッシュデータが最新のデータとなり、キャッシュデータがライトデータを上書する。

【 0 0 5 8 】

ステップ3300では、ライトデータとキャッシュデータのホスト I / F 1300で受領された時刻（受領時刻102）を比較する。ライトデータの受領時刻の方が新しい（時刻として最近）場合は、処理はステップ3310に移る。キャッシュデータの受領時刻の方が新しい場合は、処理はステップ3350に移る。本実施例において時刻の比較を行う際に、まったく同じ時刻を比較する場合の処理は含んでいない。同じ時刻を比較する場合とは、例えば記憶装置システム1070、記憶装置システム1120において計測した結果として同時に同じ領域へ書き込みが来た場合に生じる。この際の処理としては、優先記憶装置システムを決めておき、もし時刻が全く同じ場合は、優先記憶装置システム側の書き込みを優先させるなどの回避策が考えられる。

【 0 0 5 9 】

ステップ3310は、比較したキャッシュデータについてライトデータの書き込み範囲と重複する部分を削除する処理を行う。この削除処理により、キャッシュメモリ1340上の削除対象データ及びデータ一貫性保持テーブル100の対象ブロック開始アドレス104、対象サイズ105及びキャッシュデータ格納アドレス107のレコード値が変更される。

【 0 0 6 0 】

ここで例外的な処理が必要となる場合がある。この例外処理は、キャッシュデータの一部削除により、キャッシュデータの間中部分が抜き取られ、キャッシュデータが2つのレコードに分離された場合である。具体的には、削除対象のキャッシュデータの対象ブロック開始アドレス104：1000番地から対象サイズ105：200ブロックであったとする。そこにライトデータ対象ブロック開始アドレス104：1020番地から対象サイズ105：100ブロックという書き込み要求が来た場合に、キャッシュデータのデータ一貫性保持テーブル100に書かれたレコードは、ライトデータと重複部分を削除したことにより（対象ブロック開始アドレス104、対象サイズ105）が（1000、19）、（1121、79）という2つのレコードに分割される

。ここで行われる分割処理は、現在のキャッシュデータの表すレコード以降のレコードを 1 行下に下げ、テーブル管理番号101を下げたレコードに対して管理番号の変更 (+ 1) をする。この処理によって用意された 1 行下のレコード行を用いてキャッシュデータを分割し、2 つのレコードにする。

【 0 0 6 1 】

ステップ3320は、ステップ3310で変更したキャッシュデータについて、キャッシュデータがすべて削除されていないか判定する。すべて削除されている状態とは、キャッシュデータの書き込み範囲がライトデータの書き込み範囲に完全に含まれており、キャッシュデータがすべて削除されてしまう状態である。キャッシュデータの一部が削除されたときは、処理はステップ3340へ移る。キャッシュデータがすべて削除されたときは、ステップ3330の処理を行う。ステップ3330は、比較しているキャッシュデータがすべて削除されたので、データー貫性保持テーブル100の比較したキャッシュデータのレコード行を削除する。削除後、1 行うしろ以降のレコードを 1 行上にスライドさせ、移動させたレコードのテーブル管理番号101を変更 (- 1) する。

【 0 0 6 2 】

ステップ3350は、ライトデータについて比較したキャッシュデータの書き込み範囲と重複する部分を削除する処理を行う。この削除処理により、ライトデータのサイズは小さくなり、新しい対象ブロック開始アドレス104、対象サイズ105を持つことになる。もしライトデータが分割された場合は、分割されたライトデータはそれぞれ別書き込みとして書き込み処理が行われる。

【 0 0 6 3 】

ステップ3360は、ステップ3350で変更したライトデータについて、ライトデータがすべて削除されていないか判定する。ライトデータがすべて削除された状態とは、ライトデータの書き込み範囲がキャッシュデータの書き込み範囲に含まれており、ライトデータがすべて削除されてしまう状態である。ライトデータの一部が削除されたときは、処理はステップ3340へ移る。ライトデータがすべて削除されたときは、処理はステップ3370へ移る。

【 0 0 6 4 】

ステップ3340は、ステップ3310またはステップ3350で処理された重複ブロックに対応する一時ビットマップテーブル200のビット値を各々0にする。ビット値変更後に処理はステップ3250へ移る。ステップ3370は、ライトデータがすべて削除されており、書き込む内容がないのでデータ一貫性保持制御部のライトデータ書き込み処理を終了させる。

【 0 0 6 5 】

図 1 1 は、キャッシュメモリ1340上のキャッシュデータを物理ディスク1370に書き込むデータ一貫性保持制御部2040の処理（一斉書き込み処理）を示すフローチャートである。一斉書き込み処理は、タイマー1310によって6 0 秒に一回すべての記憶装置システムで一斉して起動される。ステップ3400は、データ一貫性保持テーブル100において一番昔に書かれたレコードであるテーブル管理番号101が1を変数カレントに代入する。代入後、処理はステップ3410に移る。ステップ3410では、テーブル管理番号101がカレントのレコードの受領時刻102を参照する。カレントレコードの受領時刻102が、一斉書き込み処理開始時間より3分経っていないときは処理がステップ3470へ移り、3分以上経過しているときは処理がステップ3420に移る。

【 0 0 6 6 】

ステップ3420は、カレントレコードの書き込み範囲を示すビットマップテーブル2030のビット値をすべて0に変更する。変更後処理は、ステップ3430へ移る。ステップ3430では、データ一貫性保持制御部2040は、ディスク制御部1350にカレントレコードの書き込みデータ（対象ブロック開始アドレス104、対象サイズ105、及びキャッシュデータ格納アドレス107で指定されたキャッシュメモリ1340上のデータ）を渡す。ディスク制御部1350は、物理ディスク1370にカレントレコードの書き込み内容を書き込む。ディスク制御部1350にデータを受け渡し終了後、処理はステップ3440へ移る。

【 0 0 6 7 】

ステップ3440は、ディスク制御部1350に渡したキャッシュメモリ1340上のキャッシュデータを削除する。削除後、処理はステップ3450へ移る。ステップ3450は、カレントレコードで示されるデータ一貫性保持テーブル100上のレコードの削

除を行う。カレントレコード行の削除は、データー貫性保持テーブル100のカレントレコードのテーブル管理番号101、受領時刻102、上位装置識別番号203、対象ブロック開始アドレス104、対象サイズ105、ストレージシリアル番号106及びキャッシュデータ格納アドレス107に書きこまれている情報をそれぞれ削除することである。カレントレコードの削除カレント行の削除終了後、処理はステップ3460へ移る。ステップ3460は、カレントの変数値を+1とする。カレントの示すレコードは、今回物理ディスク1370に書き込まれたレコードの次に書き込まれたレコードとなる。そして処理はステップ3410に戻り、カレントレコードが一斉書き込み処理対象レコードであるか否かを判定する。

【 0 0 6 8 】

ステップ3470は、データー貫性保持テーブル100を更新する。テーブルの先頭レコードが未使用レコードとならないようにレコードを順に上に詰め、テーブル管理番号101を上レコードから順に1、2、3、…と振りなおす。データー貫性保持テーブル100の変更後、処理は終了する。

【 0 0 6 9 】

なお第1の実施例の変形例として、ビットマップテーブル2030を設けないような実施例も可能である。その場合には、データー貫性保持制御部2040は、キャッシュメモリ1340には重複ブロックを考慮せず、書き込みデータをそのまま受領時刻の順に格納するものとする。データー貫性保持制御部2040は、ステップ3280のデーター貫性保持テーブル100への書き込みとステップ3290のキャッシュデータの書き込みを行う。重複ブロック削除処理はない。また記憶装置システムが上位装置から参照系コマンドを受けた場合のステップ3010、3020及び3040は、データー貫性保持テーブル100を最新のレコードから順に探索し、書き込みデータがキャッシュメモリ1340に存在する参照範囲についてはキャッシュメモリ1340からデータを読み取り、キャッシュメモリ1340に存在しない参照範囲についてはディスク制御部1350を介して物理ディスク1370からデータを読み取る。

(2) 第2の実施例

第1の実施例は、ライトデータの書き込み処理の際ステップ3060において、DKCI/F1320へ書き込み内容が渡される。ここでコピー先記憶装置システムの

書き込み処理の終了を待っていない。ステップ3070で、コピー先記憶装置システムの書き込み処理終了を待つか、待たないかによってコピーペアのボリューム内データの一致性が変わる。

【 0 0 7 0 】

第1の実施例では、コピー先記憶装置システムへ書き込み処理が完了する前に上位装置へ書き込み終了が通知される。しかしコピー先記憶装置システムへの書き込み処理は、リモートコピーのデータ転送時間と実際の書き込み処理をする時間を要する。この処理に要する時間の間、書き込み内容はコピー先の記憶装置システムとコピー元の記憶装置システムにおいて差異がでる。完全に2重化したボリュームをリモートコピーペアとして形成する場合には、ステップ3070でコピー先記憶装置の書き込み完了通知を待つ。コピー先記憶装置システムの書き込みにかかる時間だけ書き込まれたデータに関して一致性が失われても影響のない環境の場合には、メイン制御部2020は、コピー先記憶装置システムの書き込み完了を待たず、ホスト I / F 1300へ書き込み完了を通知する。

(3) 第3の実施例

第3の実施例は、第2の実施例に変更を加えて、リザーブ情報の伝播を実現する。第3の実施例は、ディスク領域のリザーブによってその領域の排他アクセスを行う場合に適用される。図12は、SCSIのリザーブ情報をリモートコピーの対象とする記憶装置システムに伝播させるための第3の実施例のシステム構成を示す。

【 0 0 7 1 】

図12のシステムは、排他制御部4000、データ一貫性保持制御部4020及びメイン制御4030の各プログラムを備える。これらのプログラムは、プロセッサ1380によって実行される。またメモリ上にビットマップテーブル4010を備える。その他の構成要素は、第1の実施例の通りである。

【 0 0 7 2 】

排他制御部4000は、ロック状態保持テーブル400を保持し、上位装置からのリザーブ状態を管理する。排他制御部4000は、双方向リモートコピーの対象となる記憶装置システムの各々が同じ内容のロック状態保持テーブル400を持つように

制御する。この同じ内容のロック状態保持テーブル400を持つことによって、ある記憶装置システムのボリュームを上位装置がロックした際に、コピーペアを形成している他サイトのペアボリュームもロックされている状態になる。

【0073】

ビットマップテーブル4010は、ビットマップテーブル2030と比べてビットマップが表す状態が4つになる。これによってビットマップテーブル4010は、ディスクの1ブロックに対して1ビットのビットマップではなく、2ビットでディスクの状態を保持する。ここでは通常使用される「ビットマップ」ではなくなるが、本実施例ではビットマップと呼ぶことにする。

【0074】

データ一貫性保持制御部4020は、データ一貫性保持制御部2040と比べてビットマップのとりビット値が増える（ビットマップの持つ状態が増える）ことによるビット値による処理の分岐条件に変更がある。メイン制御部4030は、メイン制御部2020と比べて、ホスト I / F 1300及びD K C I / F 1320が、参照系コマンド及び更新変更系コマンド以外にリザーブ系コマンドを受領し処理することに起因する処理の変更がある。

【0075】

図13のビットマップ値テーブル300は、ビットマップテーブル4010の持つビット値0、ビット値1、ビット値2及びビット値3が表現する状態の意味を説明するテーブルである。ビット値0は、どの上位装置にもリザーブされておらず、物理ディスク1370のデータが最新の状態を示す。ビット値1は、どの上位装置にもリザーブされておらず、キャッシュメモリ1340のデータが最新の状態を示す。ビット値2は、ある上位装置にリザーブされており、物理ディスク1370のデータが最新の状態を示す。ビット値3は、ある上位装置にリザーブされており、キャッシュメモリ1340のデータが最新の状態を示す。

【0076】

図14は、排他制御部4000が保持するロック状態保持テーブル400のデータ形式を示す。ロック状態保持テーブル400は、管理番号401、ロック開始時刻402、上位装置識別番号403、ロック対象開始アドレス404及びロック対象サイズ405と

いう項目によって構成される。ロック開始時刻402は、上位装置より記憶装置システム内のホスト I / F 1300にロック要求が受理された時刻を格納する。上位装置識別番号403は、データー貫性保持テーブル100の上位装置識別番号203と同意である。ロック開始アドレス404及び対象サイズ405は、それぞれロック対象となっているディスクのブロック番地およびブロック数を設定する。管理番号401は、テーブルの先頭より1から順に2、3、…と整数の管理番号を格納しているものとする。管理番号は1から始まり+1ずつ増加していき、最終レコードの次の管理番号は-1とする。

【0077】

図15及び図16は、メイン制御部4030の処理手順を示すフローチャートである。図15は図7に、図16は図8に変更を加えたものとなる。この変更箇所を以下に説明する。図15は、メイン制御部4030がホスト I / F 1300から I / Oを受け取ったときのフローチャートである。

【0078】

メイン制御部4030は、ホスト I / F 1300から入出力要求を受けたとき、ステップ5000において参照系のコマンド、更新変更系のコマンド及びロック系のコマンド（SCSIにおけるリザーブ、領域リザーブ、リリースなどのコマンド）を認識し、処理を分岐させる。参照系のコマンドの際は、ステップ5005へ処理が移る。更新変更系のコマンドの際は、ステップ5010に処理が移る。ロック系のコマンドの際は、ステップ5070に処理が移る。ステップ5010とステップ5005は同じ処理である。この処理は、参照および更新変更の処理を行う処理対象範囲が他の上位装置によってリザーブ系のコマンドを用いたロックをしているか否かを調べる。このロックの状態により、メイン制御部4030は、参照及び更新変更要求をした上位装置がその要求範囲について参照及び更新変更の処理が可能か否かを判定する。この判定処理の詳細については後述する。

【0079】

参照系コマンドのときは、ステップ5005の処理終了後にステップ5020の判定結果分岐へ処理が移る。更新変更系のコマンドのときは、ステップ5010の処理終了後にステップ5050の判定結果分岐へ処理が移る。参照系の要求の際、ステップ50

20において要求範囲が参照可能な場合は、処理はステップ5030へ移る。参照不可能な場合は、処理はステップ5040へ移る。更新変更系の要求の際はステップ5050において要求範囲が更新変更可能な場合は、処理はステップ5060へ移る。更新変更不可能な場合は、処理はステップ5040へ移る。ステップ5040は、参照・更新変更が不可能な際、上位装置にホスト I / F 1300を介して要求範囲が利用不可を通知する（S C S I プロトコルでは、Reservation Conflictが上位装置へ返される）。

【 0 0 8 0 】

ステップ5030は、ステップ3005、ステップ3010、ステップ3020、ステップ3040及びステップ3030の処理を行う。ここでビットマップテーブル4010においてビット値0、ビット値2のブロック参照要求は、物理ディスク1370上のデータが最新のデータという状態であり、参照データは物理ディスク1370より読み込まれるものとなる。ビット値1、ビット値3のブロック範囲の参照要求は、キャッシュメモリ1340上のデータが最新のデータという状態であり、参照データはキャッシュメモリ1340より読み込まれるものとなる。

【 0 0 8 1 】

ステップ5060は、後述するデータ一貫性保持制御部4020の処理である。ステップ5070は、ホスト I / F 1300よりロック系のコマンドがメイン制御部4030に渡された際の排他制御部4000の処理である。後述する排他制御部4000は、戻り値をメイン制御部4030に返す。戻り値の受け取り終了後、処理はステップ5080に移る。ステップ5080は、ホスト I / F 1300に排他制御部4000から受け取った戻り値を渡す。戻り値をホスト I / F 1300に送信後、ロック系のコマンドを受理した際の処理は完了となる。

【 0 0 8 2 】

図 1 6 は、メイン制御部4030がD K C I / F 1320から I / Oを受け取った時の処理手順を示すフローチャートである。ステップ5100は、更新変更系のコマンド又はロック系のコマンドを認識し、処理を分岐させる。更新変更系のコマンドの場合は、ステップ5110に処理が移る。ロック系のコマンドの場合は、ステップ5120に処理が移る。ステップ5110は、後述するデータ一貫性保持制御部4020の処理

である。ステップ5120は、D K C I / F 1320からロック系のコマンドが渡された場合の処理である。ステップ5120の処理は、排他制御部4000で行われ、メイン制御部4030は、排他制御部4000からその戻り値を受け取る。戻り値の受け取り終了後、処理はステップ5130に移る。ステップ5130は、D K C I / F 1320に排他制御部4000から受け取った戻り値を渡す。戻り値をD K C I / F 1320へ送信後、処理は完了となる。

【 0 0 8 3 】

図 1 7 及び図 1 8 は、データ一貫性保持制御部4020のフローチャートである。データ一貫性保持制御部2040からの変更点は、ビットマップテーブル4010の持つ状態が増えたことによる変更である。図 1 7 は図 9 で示したフローチャートの変更である。図 1 7 は、ステップ3200がステップ5200に、ステップ3210がステップ5210に、ステップ3270がステップ5220に処理が変更された。ステップ5200は、ライトデータの書き込み範囲についてビットマップテーブル4010の対応するビット値と比較する。この比較の結果、ライトデータの範囲のすべてのビット値が、ビット値 0 又はビット値 2 のとき（物理ディスク1370に最新のデータがあるとき）は、ステップ5220に処理が移る。ライトデータの範囲についてビットマップテーブル4010の対応するビット値がビット値 1 又はビット値 3 を含むとき（キャッシュメモリ1340上に最新のデータがあるとき）は、ステップ5210に処理が移る。

【 0 0 8 4 】

ステップ5210では、ライトデータの範囲と同じ大きさの一時ビットマップテーブル200を作成する。一時ビットマップテーブル200は、ライトデータの書き込みブロックと対応している。一時ビットマップテーブル200は、ビットマップテーブル4010において、ビット値 1 又はビット値 3 を表しているブロックにビット値 1 を代入し、ビット値 0 又はビット値 2 を表しているブロックにビット値 0 を代入する。つまりキャッシュメモリ1340上に変更データが存在するディスクのブロック位置をビット値 1 として、一時ビットマップテーブル200に格納する。

【 0 0 8 5 】

ステップ5220は、ビットマップテーブル4010のビット値を変更する処理である。処理前のビットマップテーブル4010のビット値が 0 又は 1 の場合は 1 に変更し

、ビット値が2又は3の場合は3に変更する。

【0086】

図18は、キャッシュメモリ1340上のキャッシュデータを物理ディスク1370に書き込む処理（一斉書き込み処理）のステップ3420を変更し、ステップ5300にしたものである。ステップ5300は、ビットマップテーブル4010のビット値の変更を行う処理であるが、変更対象のビットマップテーブル4010のビット値1の場合はビット値0へ、ビット値3の場合はビット値2へ変更する処理を行う。これは一斉書き込み処理によってキャッシュメモリ1340上のキャッシュデータがなくなることによるビットマップビット値の変更処理である。

【0087】

図19、図20及び図21は、排他制御部4000に関する処理のフローチャートである。図19は、上位装置の書き込み要求が来た際に、書き込み範囲に書き込み処理が可能か、すなわち書き込み範囲が書き込み要求を出した上位装置以外によって既にロックされていないか、判定する処理を示している。

【0088】

ステップ5400は、処理要求範囲（参照、更新変更要求されている範囲）に対応するビットマップテーブル4010を参照することにより、上位装置によって要求範囲がロックされていないか判定する。要求範囲がすべてビット値0又は1の場合は、処理要求範囲はどの上位装置にもロックされておらず、排他制御部4000は、利用可能を戻り値に設定する。処理要求範囲にビット値2又はビット値3が含まれる場合は、ステップ5410に処理が移る。ステップ5410では、処理要求範囲をロックしている上位装置が処理要求をしている上位装置であるかロック状態保持テーブル400のレコードを参照して判定する。この判定は、処理要求範囲に対応するビットマップテーブル4010においてビット値2又はビット値3のロック状態保持テーブル400のレコードをすべて検索し、検索されたレコードの上位装置識別番号403が今回処理要求をしている上位装置であるか否かを判定する。処理要求をしている上位装置によるロックの際には、今回のライトデータは処理可能であり、利用可能を戻り値に設定する。処理要求をしている上位装置以外の処理要求の際は、利用不可を戻り値に設定する。

【0089】

図20は、ホストI/F1300が上位装置からディスクのロック要求を受理した際の排他制御部4000のフローチャートである。ステップ5500は、ロック要求範囲（ロックを要求しているディスクのブロック範囲）が既に他の上位装置によりロックされていないか、ビットマップテーブル4010のビット値を参照して判定する。ロック要求範囲に対応するビット値がすべて0又は1のとき（どの上位装置もロックしていない状態）、処理はステップ5530に移る。ロック要求範囲に対応するビット値に2又は3を含む場合は、ロック要求範囲がある上位装置によって既にロックされていることになる。この際、ステップ5505でロック指定範囲のリザーブが今回ロックを要求している上位装置のリザーブ状態であるか判定する。ロックをしている上位装置からのロック要求である際は、ロック完了の戻り値をセットする。

【0090】

別の上位装置のロック状態の際は本来、ロック失敗を戻り値に設定する。しかしリモートコピー環境下においてコピーペアの他のサイトがロック解消処理を実行中であり、自サイトへその処理結果が届く間にロック要求を受理したということもあり得る。よってステップ5510の処理によってコピー先記憶装置システムにロック要求を送信する。この戻り値をステップ5520で判定し、ロック失敗が相手サイトの記憶装置システムより戻ってきた際は、ロック失敗を戻り値に設定する。処理5520よりロック完了を受信した際は、もう一度ステップ5500に処理を移し、ロック処理を最初から行う。

【0091】

ステップ5505、ステップ5520は、まず自ホスト内ビットマップテーブル4010参照による他の上位装置によるロックされている状態であるかどうかという判定する。次に他サイトへのロック要求の戻り値を見る。この処理は、自サイトのビットマップを判定しロックが不可能という判定後、上位装置にロック失敗を伝達するため、上位装置が再びロック要求を出す時間より早く、効率的である。他サイトにロック要求を送信してそのロック要求を待つ処理は、リモートコピー環境下のリザーブ情報伝播において伝播にかかるタイムラグに対応するために必要な処

理となる。

【 0 0 9 2 】

ここで具体的にステップ5520の処理が必要となる環境について図 2 2 及び図 2 3 のコンピューターシステム6000を用いて説明する。図 2 2 は、このコンピューターシステムの構成図である。このシステムは、サイト1110とサイト1120のホスト A 6010とホスト B 6020がクラスタ環境を構築している。このクラスタは論理ボリュームをクラスタのリソースとして運用していると仮定する。サイト1110内には論理ボリューム A 6040があり、サイト1120内には論理ボリューム B 6050がある。それぞれの論理ボリュームは、双方向リモートコピーによってコピーペアが形成されている。両ホストは、これら複数の論理ボリュームを同一の論理ボリュームとして扱っているものとする。ホスト A とホスト B は、I P (Internet protocol) ネットワーク6030を用いて通信するものとする。

【 0 0 9 3 】

次に図 2 3 の状態遷移と処理の例について説明する。このクラスタでは S C S I のリザーブコマンドによってディスクの排他制御を行うことがある。例えばマイクロソフト社のクラスタサーバーなどである。ここでサイト1110のホスト A 6010が論理ボリューム A 6040をオフラインのリソースとして運用をするとき、ホスト A 6010は、論理ボリューム A 6040をリザーブしている。次にホスト A 6010の業務をホスト B 6020へ、使用していた論理ボリューム A 6040をリモートコピーペアである論理ボリューム B 6050に移行（フェイルオーバー）するものとする。ホスト A 6010は、論理ボリューム A 6040をリリースし、ホスト B 6020は、論理ボリューム B 6050をリザーブし、ホスト B 6020は論理ボリューム B 6050を用いて運用を行う。ホスト A 6010とホスト B 6020は I P ネットワーク6030を用いてフェイルオーバーを連絡して処理の移行を円滑に行う。

【 0 0 9 4 】

この際、ホスト A 6010のリリースとホスト B 6020のリザーブがほぼ同時に実行されると、ステップ5510及びステップ5520の処理がない場合は、ホスト B 6020は、論理ボリューム B 6050をリザーブすることができない。すなわち論理ボリューム A 6040がリリースされているにも関わらず、論理ボリューム B 6050がリリース

されていないため、ホスト B 6020 が論理ボリューム B 6050 をリザーブできない。こうしてリリースした論理ボリュームをリザーブできないことによって、クラスターサーバーの動作に影響する可能性が生じる。よってリザーブ要求の結果をすぐリザーブ要求ホストへ返すのではなく、ステップ 5510、5520 のように相手サイトへのリザーブ要求の返信を待ち、ホストに返す処理が必要となる。

【 0 0 9 5 】

ステップ 5530 は、ロック対象領域に相当するビットマップテーブル 4010 のビット値を 2 又は 3 にする。具体的には、ビットマップテーブル 4010 の変更対象ビットマップのビット値が 0 の場合は 2 へ、ビット値が 1 の場合は 3 へ変更する処理を行う。ビット値変更後、処理はステップ 5540 に移る。ステップ 5540 は、ロック状態保持テーブル 400 の管理番号 401 が - 1 のレコード位置に、ロック要求のロック開始時刻 402、上位装置識別番号 403、ロック対象開始アドレス 404 及びロック対象サイズ 405 のレコードを登録する。その管理番号 401 は、一つ上のレコードの管理番号 401 に + 1 した整数が設定される。次に排他制御部 4000 は、登録したレコードの一つ下の空きレコードの管理番号 401 に - 1 を設定する。このレコードの書き込み処理終了後、ステップ 5550 に処理を移す。ステップ 5550 は、相手サイト記憶装置システムにロック要求を送信する。ここで送信方法について特に記載しないが、S C S I のライトコマンドを拡張したベンダ固有コマンドを用いてもよい。

【 0 0 9 6 】

ステップ 5510 及びステップ 5550 の処理では、ロック状態保持テーブル 400 のレコード項目及びロック要求時の戻り値は、リモートコピーライン 1060 を介して D K C I / F 1320 に渡される。ステップ 5550 は、相手サイトからの戻り値を判定する。ロック完了が戻り値の際は、ロック完了を戻り値にセットする。ロック失敗が戻り値の際は、ロック失敗を戻り値にセットする。

【 0 0 9 7 】

ステップ 5550 及びステップ 5560 の処理の結果、相手サイトからロック失敗が戻ってくる場合について説明する。この場合は、記憶装置システム 1070 と記憶装置システム 1080 がほぼ同時に上位装置よりロック要求を受け取った時である。この

場合に、サイト1110のロック要求が失敗したときは、サイト1120のロックが成功する。そこで記憶装置システム1070はステップ5530及びステップ5540で行った処理をキャンセルしなくてはならないが、このキャンセル処理は、後述するようにロックが成功した記憶装置システム1080の記憶装置システム1070へのロック形成処理で行われる（図 2 1 のステップ5650の処理参照）。

【 0 0 9 8 】

図 2 1 は、D K C I / F 1320が他のサイトよりロック要求を受理した際の排他制御部4000の処理手順を示す。ステップ5600は、ビットマップテーブル4010中のロック要求範囲に対応するビットマップのビット値を判定する。ロック要求範囲がすべてビット値 0 又は 1 の場合は、ステップ5530に移る。ロック要求範囲にビット値 2 又は 3 が含まれる場合は、ステップ5610に処理が移る。ステップ5610は、ロック要求領域に対応するサイズの一時ビットマップテーブル200を作成し、ビットマップテーブル4010の対応するビット値 2、ビット値 3 に注目して一時ビットマップテーブル200にビットマップ情報を書き込む。

【 0 0 9 9 】

ビットマップ情報の書き込み時、ビットマップテーブル4010でビット値 2、ビット値 3 の際にはビット値 1 を、ビットマップテーブル4010でビット値 0、ビット値 1 の際にはビット値 0 を一時ビットマップテーブル200に書き込む。このように一時ビットマップテーブル200は、該当ブロックがロックされているか否かのみに注目し、重複されてロックされているブロックに対応するビットにはビット値 1 を、ロックされていないブロックに対応するビットにはビット値 0 を持つビット値になる。

【 0 1 0 0 】

一時ビットマップテーブル200作成後、ステップ5620に処理を移動する。ステップ5620は、ロック状態保持テーブル400の一番新しく書き込まれたレコード（管理番号401が- 1 の一つ上のレコード）の管理番号401を変数カレントに代入する。代入後処理はステップ5630に移る。

【 0 1 0 1 】

ステップ5630は、一時ビットマップテーブル200のビット値1の範囲とカレント

レコードのロック範囲において重複してロックしている部分があるか否か判定する。重複する場合は、一時ビットマップテーブル200の重複部分のビット値を0にし、ステップ5640の処理に移る。重複がない場合は、ステップ5680の処理に移る。ステップ5680は、変数カレントを-1にし、処理をステップ5690に処理を移す。ステップ5690は、今回のロック要求時刻とカレントレコードのロック開始時刻402を比較する。カレントレコードのロック開始時刻402のほう古い（時刻が前）場合は、当該ブロック範囲には今回のロック要求時刻よりも昔にロックされているブロック範囲があるので、ロック失敗を戻り値に設定する、カレントレコードのロック開始時刻402のほう新しい（時刻が後）場合は、ステップ5630の処理に移る。まったく同じ時刻を比較する場合の処理は、ステップ3300の場合と同様である。

【0 1 0 2】

ステップ5640では、一時ビットマップテーブル200のビット値がすべて0になっているか判定する。すべてビット値が0となっている際は、当該ブロック範囲には今回のロック要求時刻よりも新しいロック要求しか来ていないこととなり、処理はステップ5650へ移る。ビット値に1を含む際は、処理はステップ5680へ移る。

【0 1 0 3】

ステップ5650は、ロック要求時刻よりあとにロックされ、ロック要求範囲と重複したレコードについて、ロック状態保持テーブル400からそのレコードを削除する。削除後、ロック状態保持テーブル400のレコードを上から順に空のレコードがないように整列し、管理番号を振りなおす。ビットマップテーブル4010において削除するレコード範囲を表すビット値を、ビット値3であった場合にはビット値1へ、ビット値2であった場合にはビット値0へ変更する。次にステップ5530及びステップ5540を順に実行しロック完了を戻り値にセットする。

【0 1 0 4】

図2 4及び図2 5は、ホスト I / F 1300及びD K C I / F 1320がロック解消要求（S C S I のリリースコマンドなど）を受理した際の処理となる。図2 4は、ホスト I / F 1300がロック解消要求を受理した際の処理を示し、図2 5は、D K

C I / F 1320が他のサイトよりロック解消要求を受理した際の処理を示す。

【 0 1 0 5 】

ステップ5700は、ロック解消要求範囲に対応するビットマップテーブル4010のビット値を、処理前のビット値が2のときは0へ、ビット値が3のときは1へ変更を行う。変更後処理はステップ5710へ移る。ステップ5710は、ロック解消のロック状態保持テーブル400のレコードを削除する。削除後ロック解消されたレコードより下に書かれたレコードを一つ上に移動させ、移動させた後移動させたレコードの管理番号401を振りなおす。管理番号401は移動させたレコードの管理番号を-1すればよい。ステップ5720は、ロック解消要求をコピーペアの各記憶装置システムのD K C I / F 1320へ送信する処理である。図 2 4 及び図 2 5 のいずれの場合もロック解消処理終了後、ロック解除を戻り値に設定する。

【 0 1 0 6 】

図 2 4 及び図 2 5 のロック解消処理は、S C S I コマンドのリリース、リセットなど各種コマンドを分けていない。本実施例は、すべての上位装置よりロック解消が来た際に、ロック解消要求範囲についてはロックが解消されるという処理になっている。図 2 4 及び図 2 5 の処理において、上位装置の識別判定処理などを入れると、更に厳密にS C S I プロトコルに準じたロック解消をするリモートコピーペア環境に適応できる。

(4) 第4の実施例

実施例 3 は、実施例 2 の双方向リモートコピーを用いた際のリザーブ情報の伝播を実現した。実施例 4 は、実施例 1 の双方向リモートコピーを用いた際のリザーブ情報の伝播の実現方法を示す。

【 0 1 0 7 】

実施例 1 の双方向リモートコピーは、書き込み内容がD K C I / F 1320へ渡される。ここでコピー先記憶装置システムの書き込み処理の終了を待っていない。この状態ではリザーブ処理中に書き込みが行われる可能性がある。リザーブ処理中の書き込み内容に関して、リザーブ要求時刻後に書き込まれた内容を、そのまま書き込み内容として扱うか、書き込み内容はなかったものとして削除するかを選択する必要がある。この処理を図 2 0 及び図 2 1 のロック完了を戻り値に設定

する前に入れることとなる。

【0108】

ここで書き込み内容を削除し、書き込みが行われなかったことにするための処理を示す。ロック完了を戻り値に設定する前に、データー貫性テーブルのロック対象範囲にロック開始時刻以降に書き込まれたライトデータがあるか否かを検索をし、書き込みデータが存在するとき、この書き込みデータは書き込みが行われなかったこととなり、データー貫性保持テーブル100内のレコード及びキャッシュメモリ1340上のキャッシュデータが削除される。この変更により、実施例1の双方向リモートコピーを用いた際のリザーブ情報の伝播を実現できる。

(5) 第5の実施例

第5の実施例は、第3の実施例を変更し、リザーブ情報の伝播を別の手段で実現する。第5の実施例は、記憶装置システム内の排他制御部4000がロック状態保持テーブル400を持たず、物理ディスク1370に直接リザーブ系コマンドを送信して、SCSIのプロトコルを用いてロック状態を管理する。このときコピーペア先の記憶装置システムへはリモートコピーライン1060を介してコピー先の物理ディスク1370も上位装置のSCSI IDを用いてリザーブを行う。このリザーブ処理は、ディスク制御部1350がサード・パーティ・リザーブを用いて、物理ディスク1370をロック要求上位装置からのリザーブとしてロックする。上位装置よりの参照、変更更新などの要求の際も、まず物理ディスク1370を用いて利用可能状態か判定する。

【0109】

この実施例は、SCSIプロトコルをそのまま用いるため、リザーブ属性などのリザーブ情報の伝播が正確に実施できる。ここでリザーブ状態はディスクを直接SCSIコマンドによって管理されるが、データはデーター貫性保持制御部4020によって管理される。よって物理ディスク1370は正確なりザーブ状態を持つが、最新のデータを持たないという状態も出てくる。

【0110】

【発明の効果】

本発明によれば、複数の記憶装置システム間でコピーペアを構成するとき、コ

ピーペアを構成するボリュームは双方向にコピーを行うことができる。各上位装置は、コピーペアを形成するどのボリュームにも自由に書き込むことができる。また双方向コピーの下で記憶装置システム間でリザーブ状態を伝播できるようになる。

【図面の簡単な説明】

【図 1】

実施形態のコンピューターシステムの構成図である。

【図 2】

実施形態の双方向リモートコピーを説明する図である。

【図 3】

実施形態の記憶装置システムのハードウェア構成図である。

【図 4】

第 1 の実施例の記憶装置システムのソフトウェア構成図である。

【図 5】

第 1 の実施例のデータ一貫性保持テーブルの構成図である。

【図 6】

第 1 の実施例の一時ビットマップテーブルの例を示す図である。

【図 7】

第 1 の実施例のホスト I / F より受けた I / O を処理する処理手順を示すフローチャートである。

【図 8】

第 1 の実施例の D K C I / F より受けた I / O を処理する処理手順を示すフローチャートである。

【図 9】

第 1 の実施例のデータ一貫性保持制御部の書き込み要求時のフローチャートである。

【図 1 0】

第 1 の実施例のデータ一貫性保持制御部の重複ブロック削除処理のフローチャートである。

【図 1 1】

第 1 の実施例のデータ一貫性保持制御部の一斉書き込み処理のフローチャートである。

【図 1 2】

第 3 の実施例の記憶装置システムのソフトウェア構成図である。

【図 1 3】

第 3 の実施例のビットマップ値を説明する図である。

【図 1 4】

第 3 の実施例のロック状態保持テーブルの例を示す図である。

【図 1 5】

第 3 の実施例のホスト I / F より受けた I / O を処理する際のフローチャートである。

【図 1 6】

第 3 の実施例の D K C I / F より受けた I / O を処理する処理手順を示すフローチャートである。

【図 1 7】

第 3 の実施例のデータ一貫性保持制御部の書き込み要求時のフローチャートである。

【図 1 8】

第 3 の実施例のデータ一貫性保持制御部の一斉書き込み処理のフローチャートである。

【図 1 9】

第 3 の実施例の排他制御部の上位装置利用可・不可判定処理のフローチャートである。

【図 2 0】

第 3 の実施例の排他制御部のホスト I / F より受けたロック形成要求によるロック形成処理のフローチャートである。

【図 2 1】

第 3 の実施例の排他制御部の D K C I / F より受けたロック形成要求によるロ

ック形成処理のフローチャートである。

【図 2 2】

第 3 の実施例を適用するコンピューターシステムの構成図である。

【図 2 3】

図 2 2 のシステムの処理の一例に関する処理シーケンス図である。

【図 2 4】

第 3 の実施例の排他制御部のホスト I / F より受けたロック解消要求によるロック解消処理のフローチャートである。

【図 2 5】

第 3 の実施例の排他制御部の D K C I / F より受けたロック解消要求によるロック解消処理のフローチャートである。

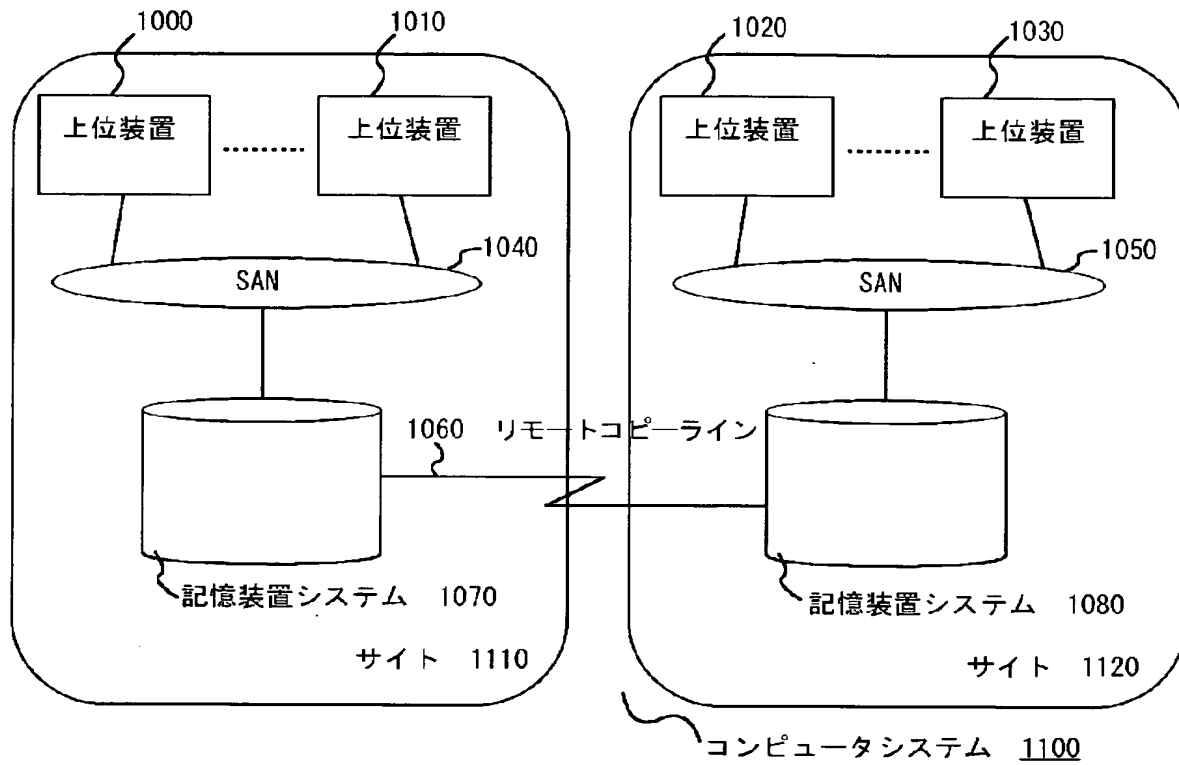
【符号の説明】

1000：上位装置、1070：記憶装置システム、1080：記憶装置システム、1370：物理ディスク、2040, 4020：データ一貫性保持制御部、4000：排他制御部。

【書類名】 図面

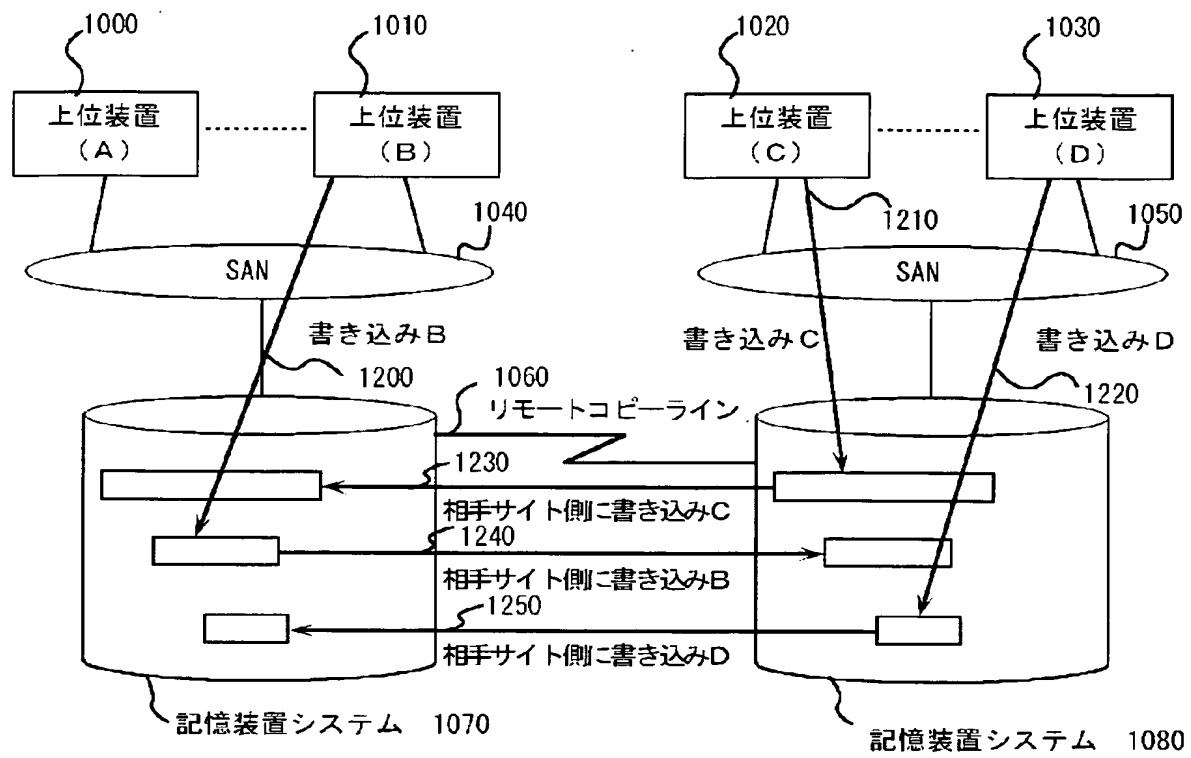
【図 1】

図 1



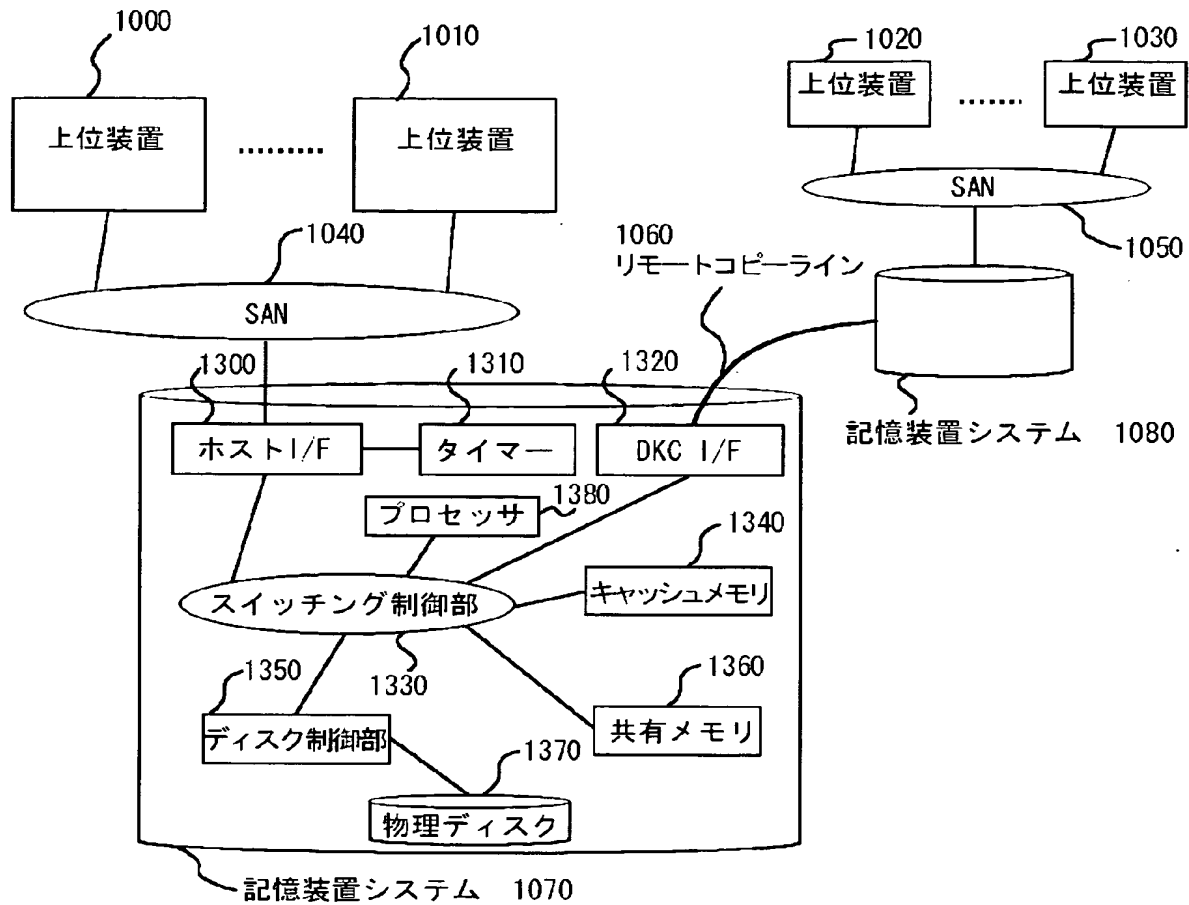
【図 2】

図 2



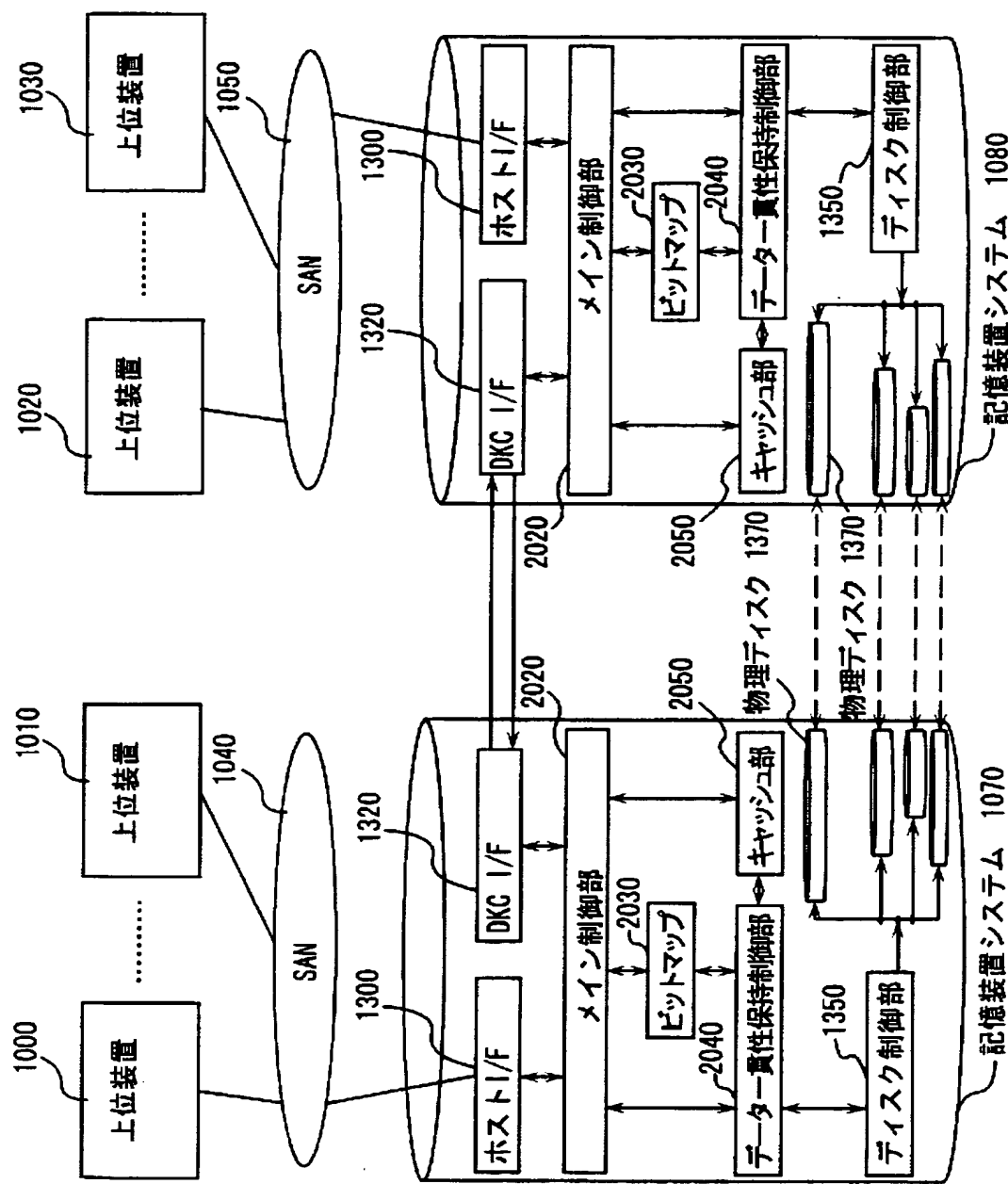
【図 3】

図 3



【図 4】

図 4



【図 5】

図 5

100 データ一貫性保持テーブル

101 テーブル管理 番号	102 受領時刻	103 上位装置識別 番号	104 対象ブロック 開始アドレス	105 対象サイズ	106 ストレージ シリアル番号	107 キャッシュデータ 格納アドレス
1	00:00:00	-----	-----	-----	-----	-----
2	00:01:30	-----	-----	-----	-----	-----
3	00:01:40	-----	-----	-----	-----	-----
⋮						
100	00:03:10	-----	-----	-----	-----	-----
-1						

【図 6】

図 6

2030 ビットマップテーブル

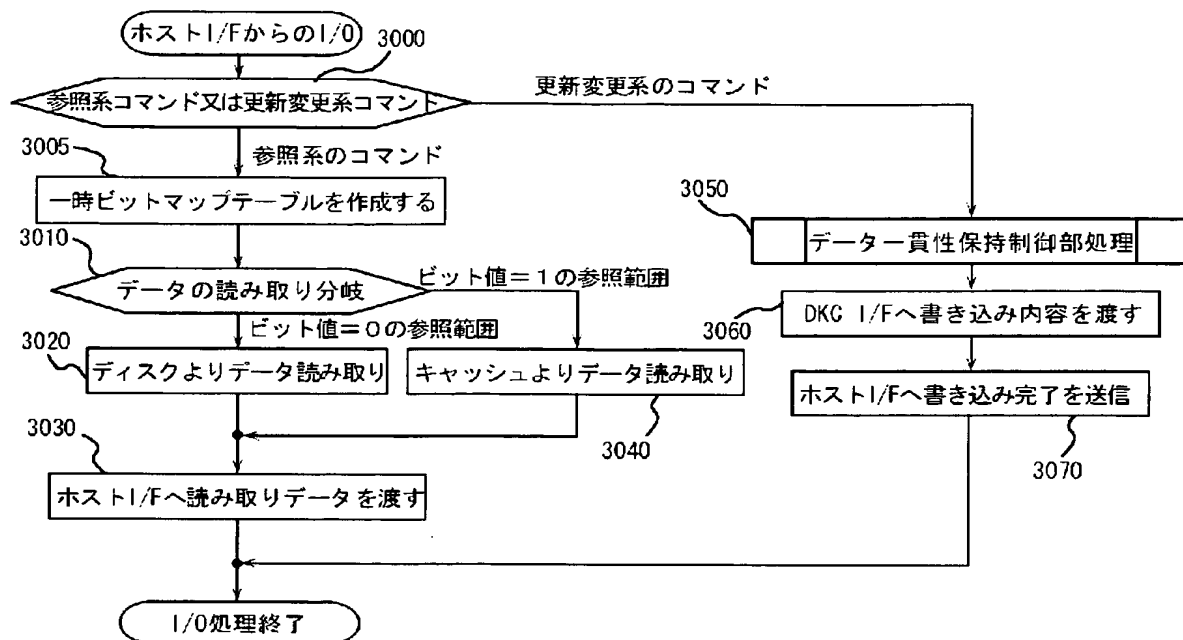
1	0	0	0	0	1	1	0	1	0
1	0	0	0	0	0	0	1	1	1
1	1	1	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	1	1	1
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
1	0	0	0	1	1	1	0	0	0
0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0

200 一時ビットマップテーブル

1	1	1	0	0	0	0
---	---	---	---	---	---	---

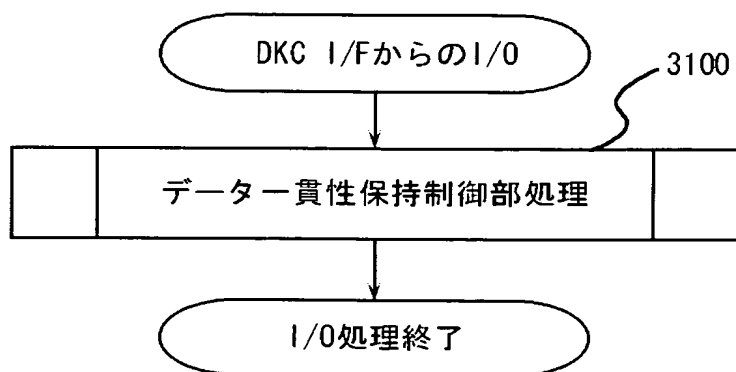
【図 7】

図 7



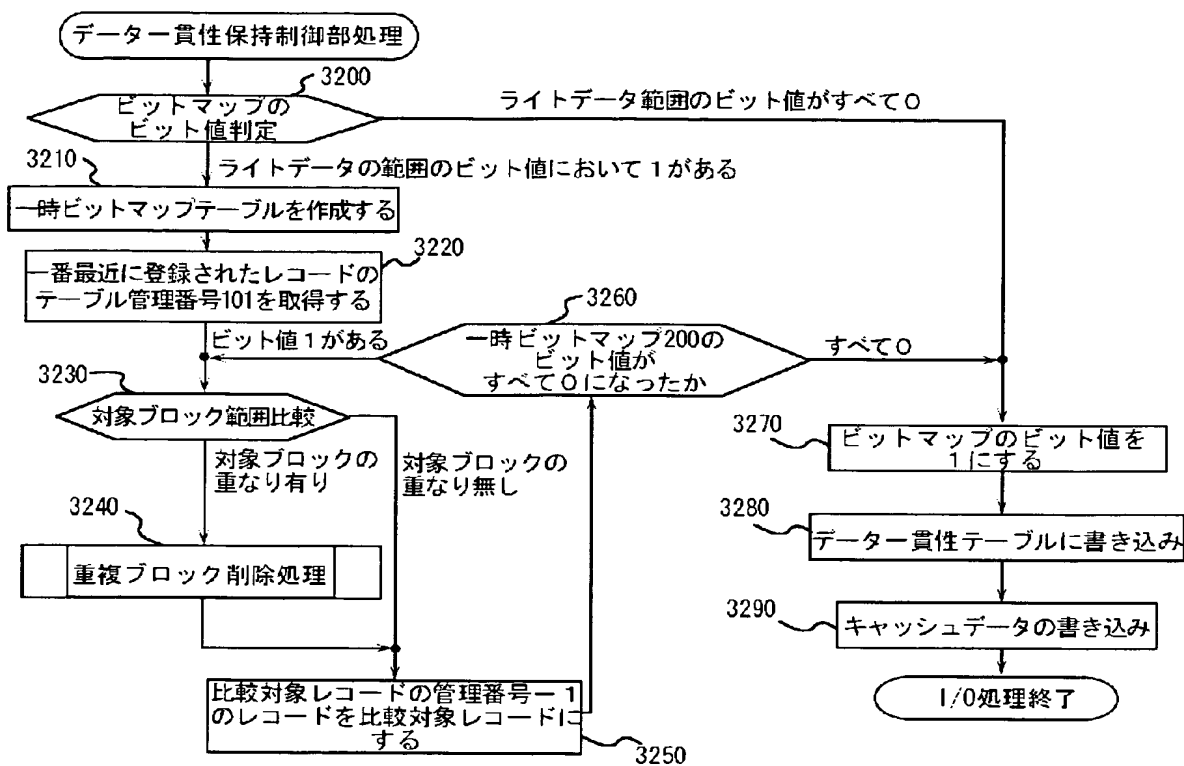
【図 8】

図 8

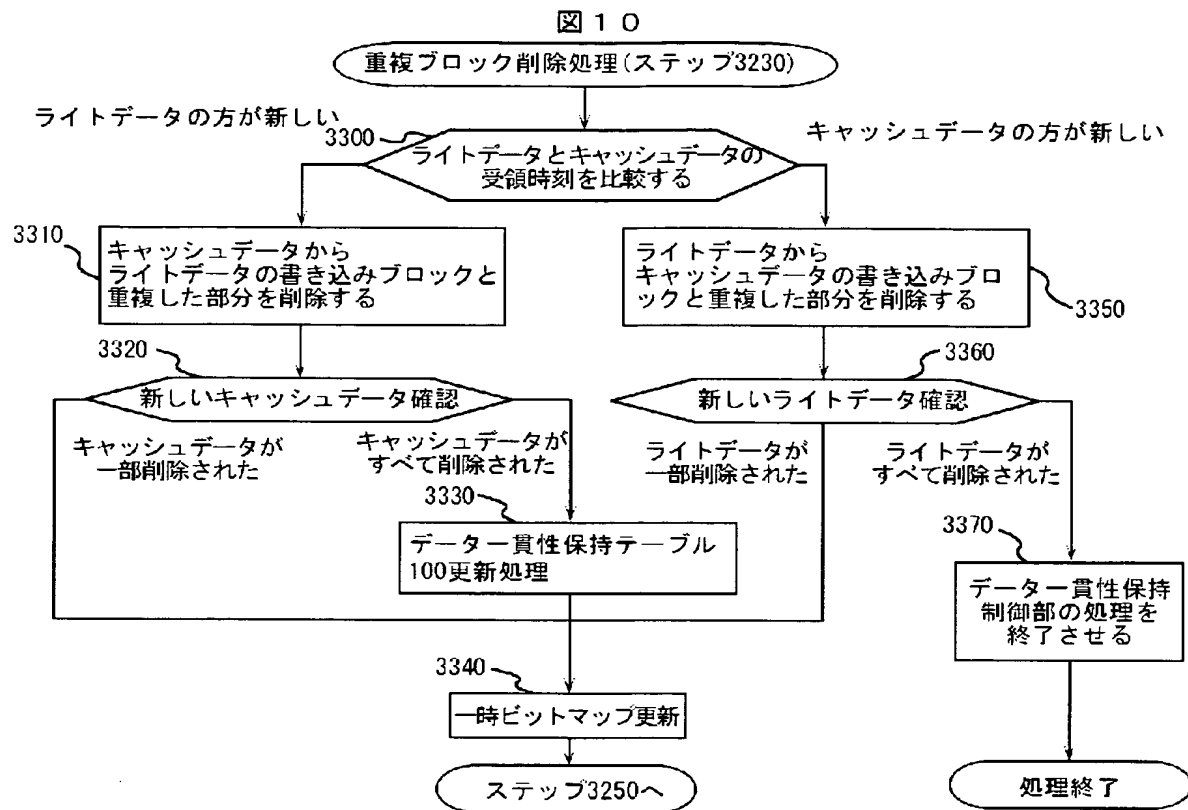


【図 9】

図 9

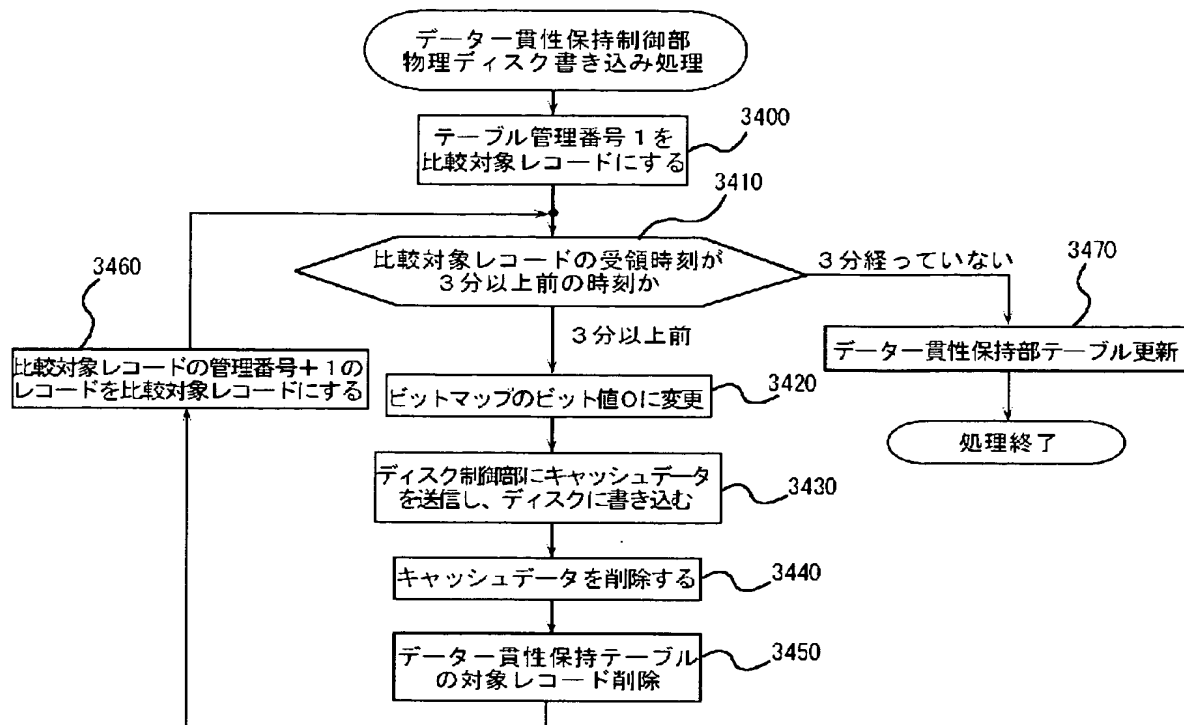


【図 10】

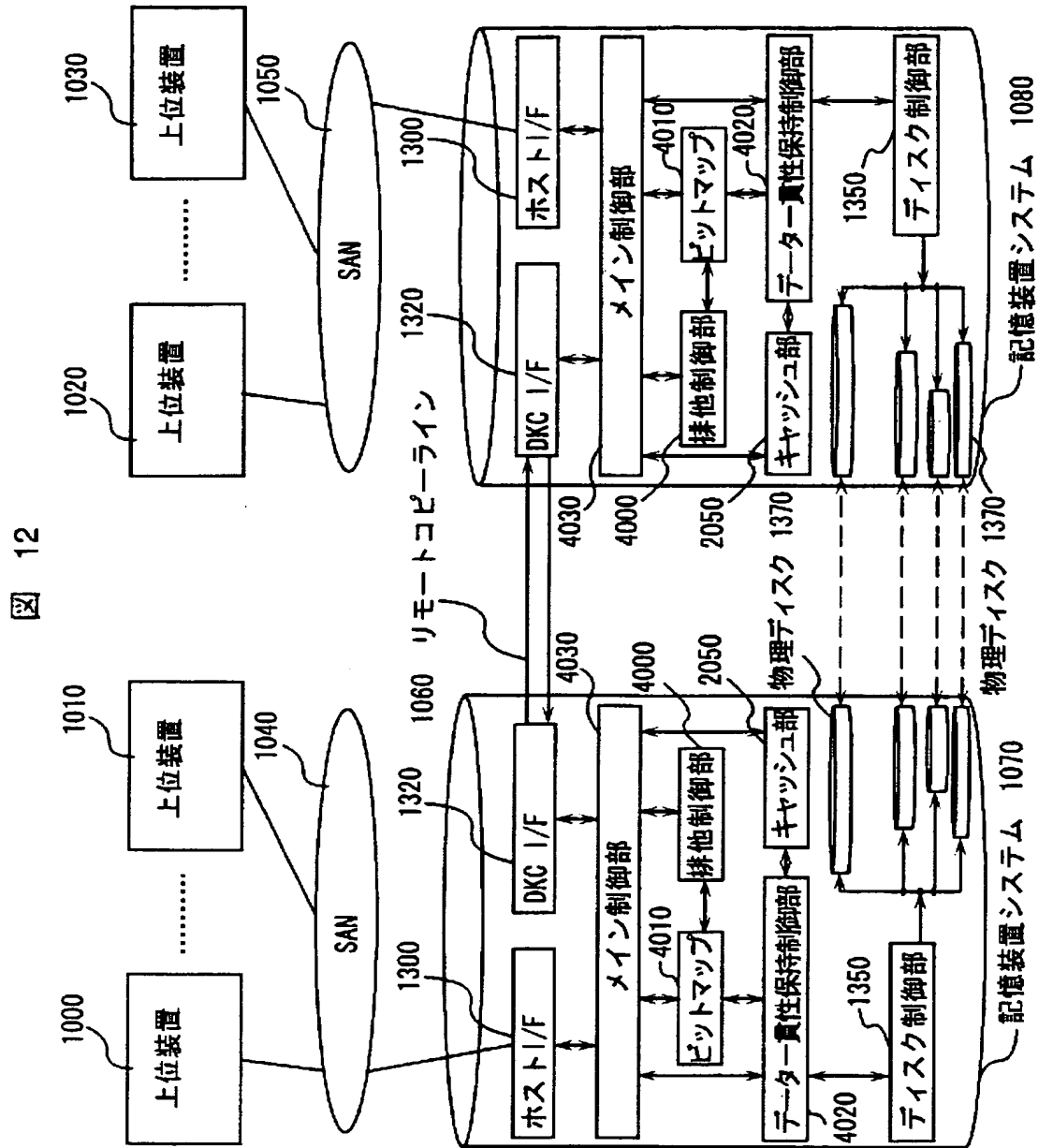


【図 11】

図 11



【図 12】



【図 13】

図 13

300 ビットマップ値テーブル

ビット値	説明
0	ディスクはリザーブされておらず ディスクが最新の状態
1	ディスクはリザーブされておらず ディスクに更新されていないデータがキャッシュ上に存在する
2	ディスクがある上位装置にリザーブされ、 ディスクが最新の状態
3	ディスクがある上位装置にリザーブされ、 ディスクに更新されていないデータがキャッシュ上に存在する。

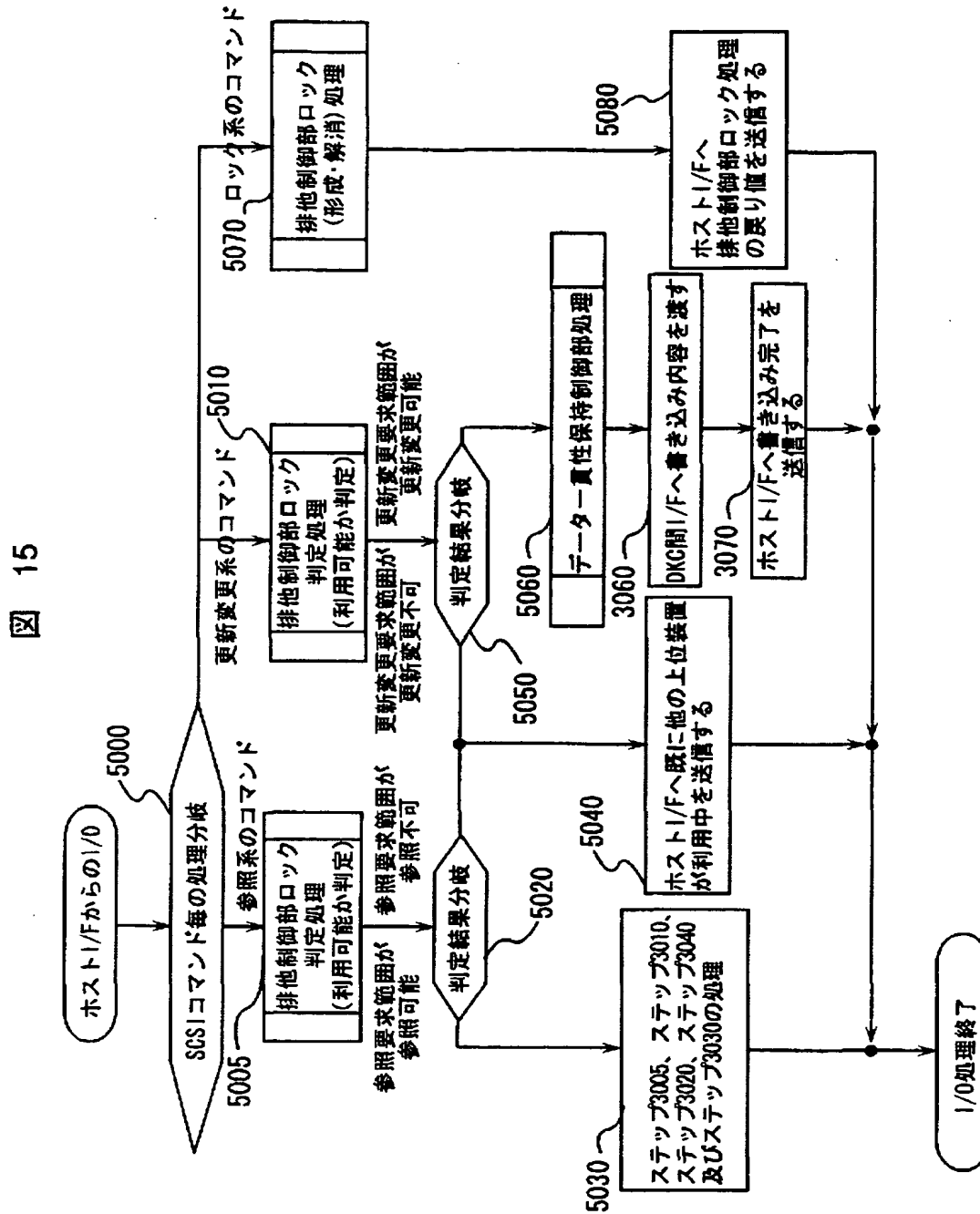
【図 14】

図 14

400 ロック状態保持テーブル

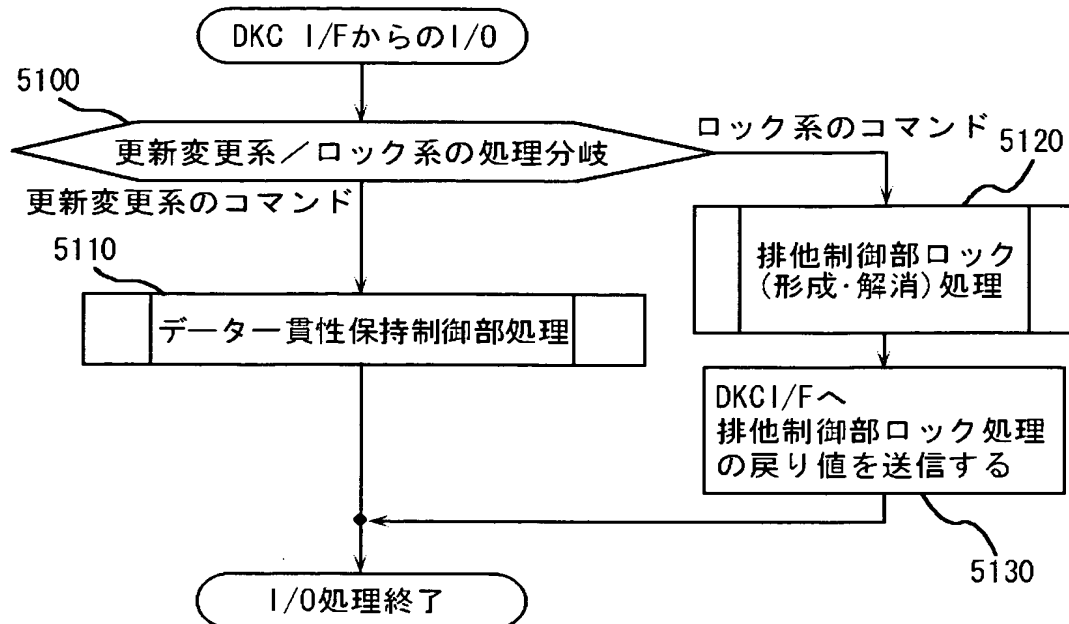
管理番号	ロック開始時刻	上位装置識別番号	ロック対象開始アドレス	ロック対象サイズ
⋮				

【図 15】



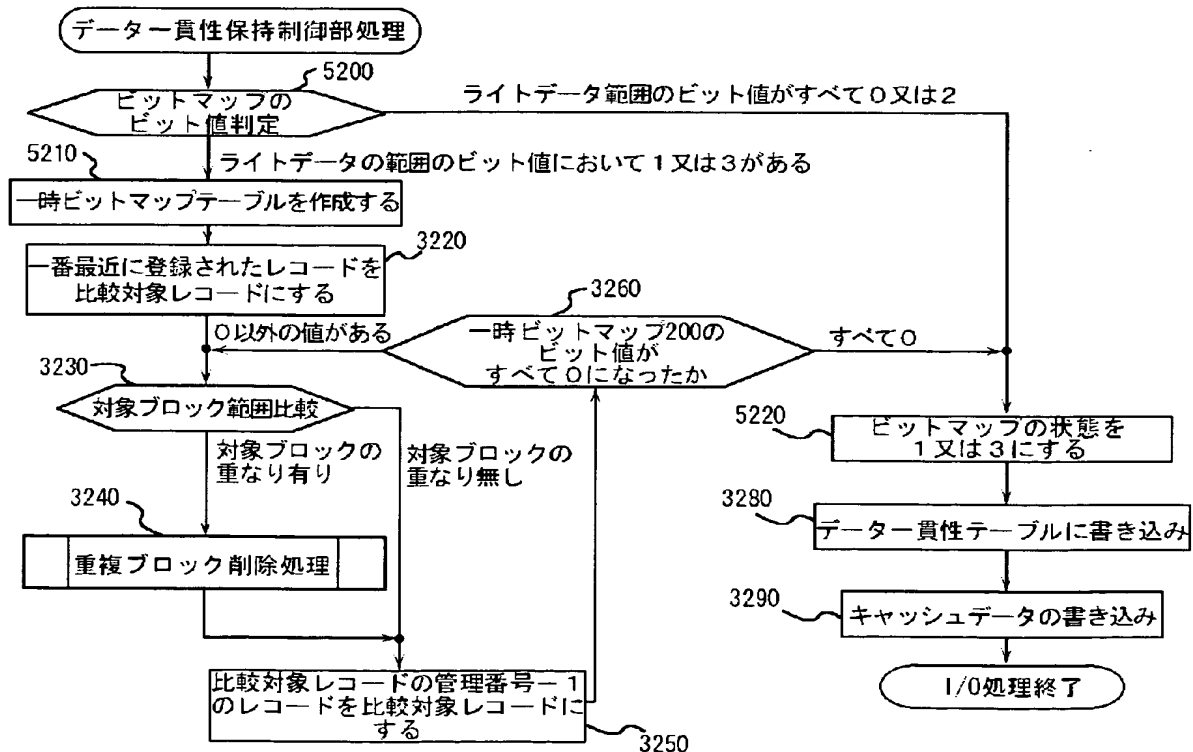
【図 16】

図 16



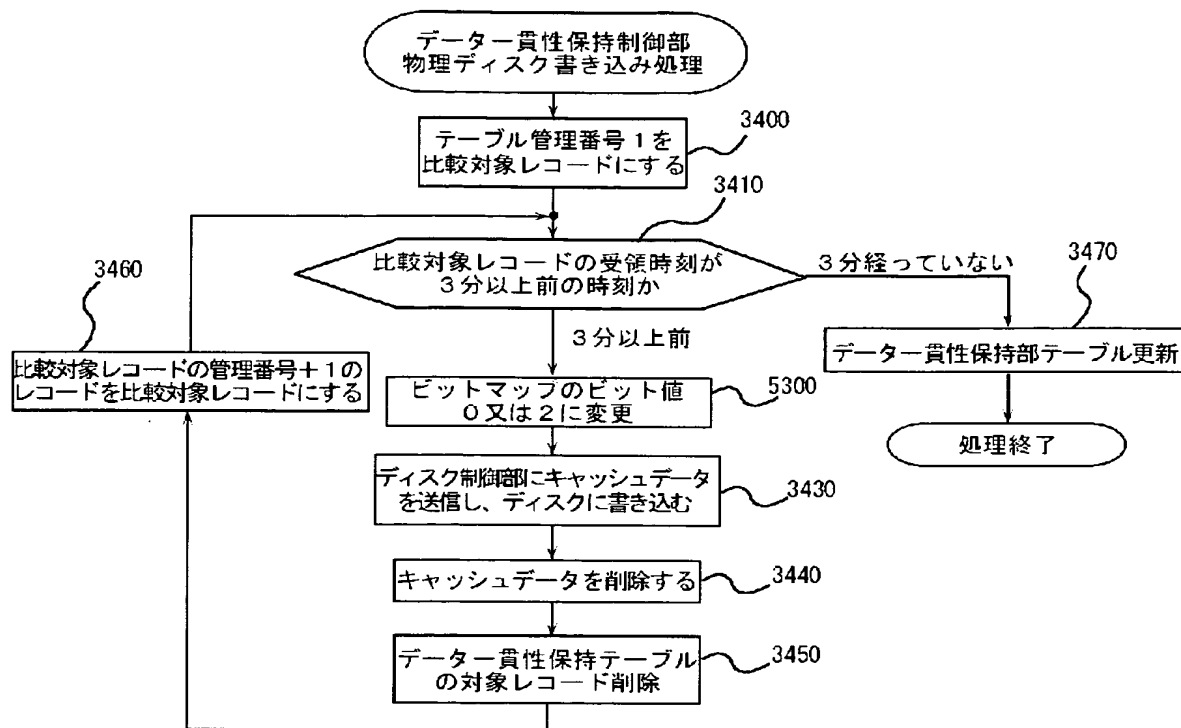
【図 17】

図 17



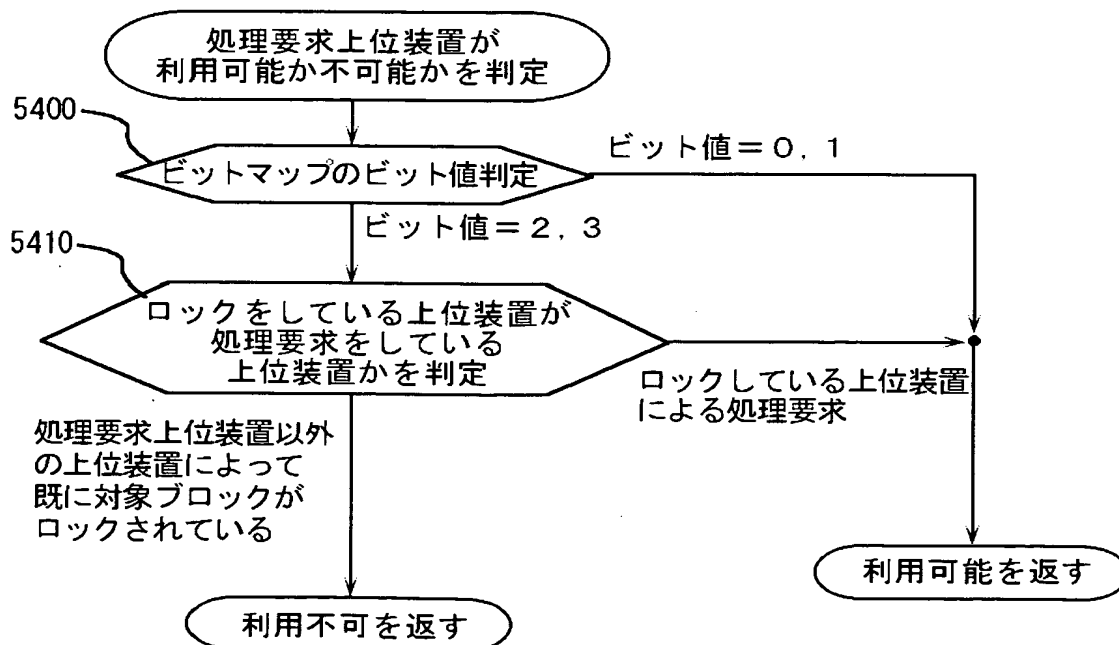
【図 18】

図 18



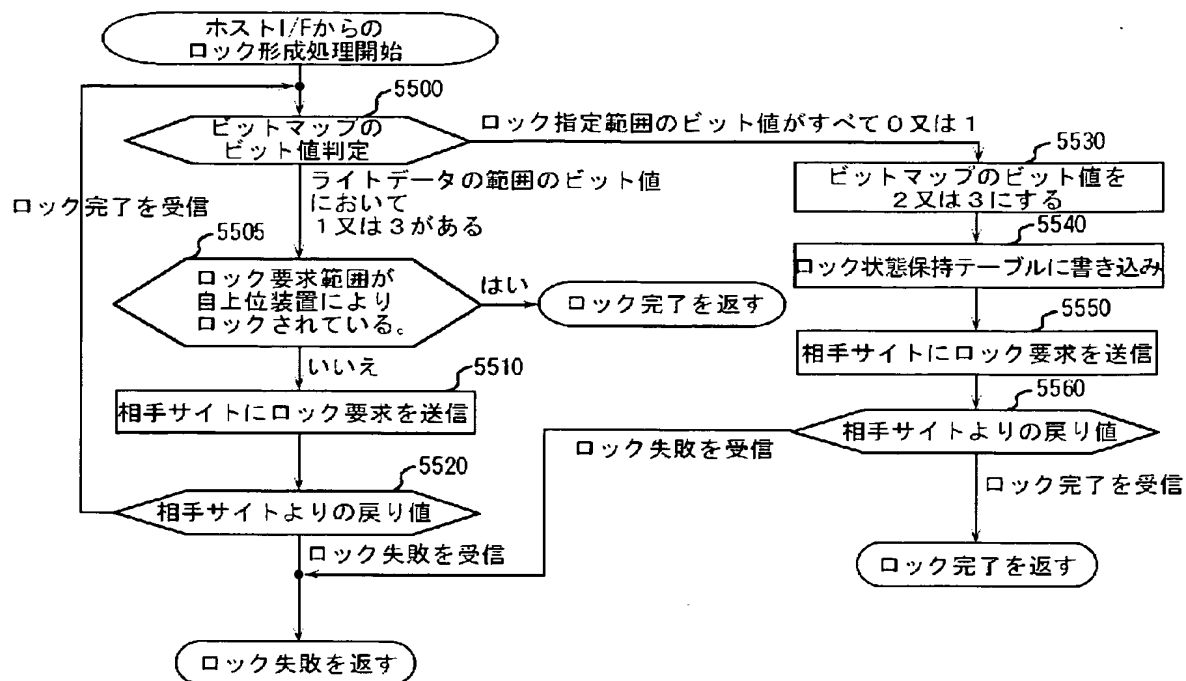
【図 19】

図 19



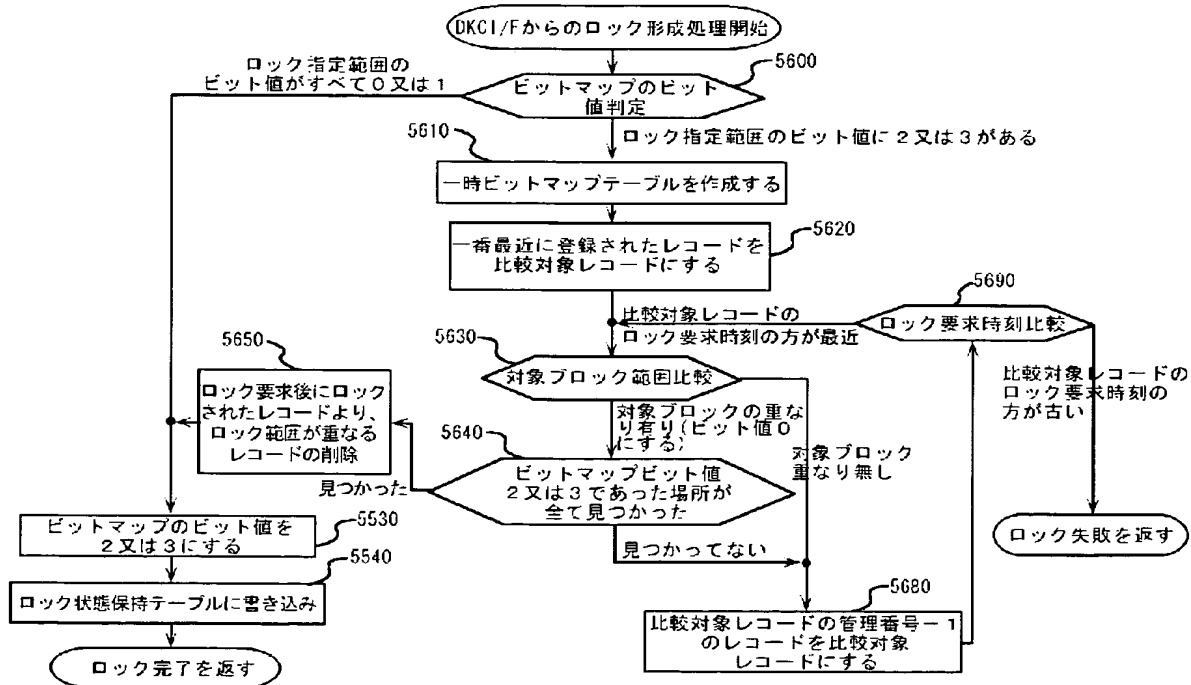
【図 20】

図 20



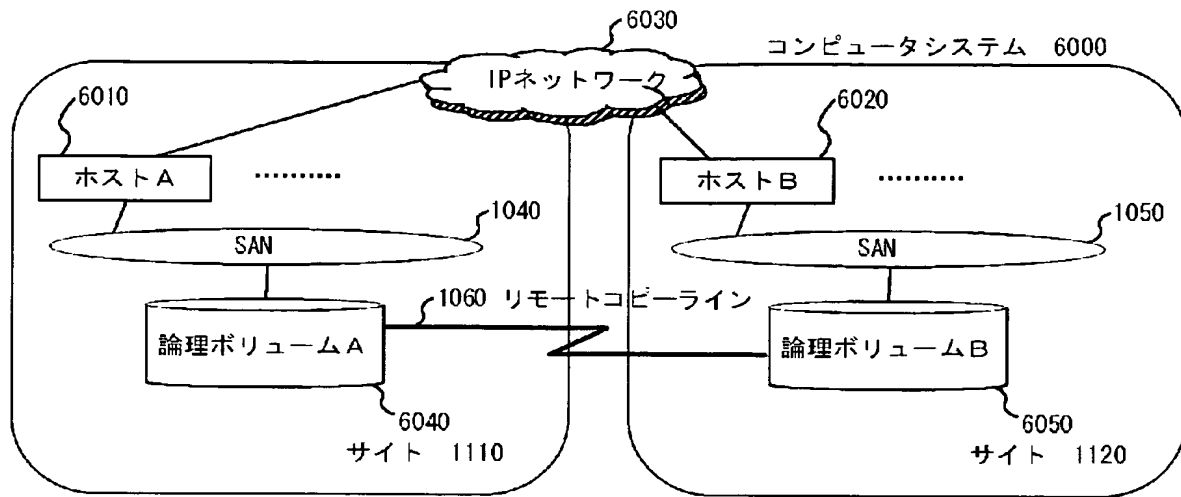
【図 21】

図 21



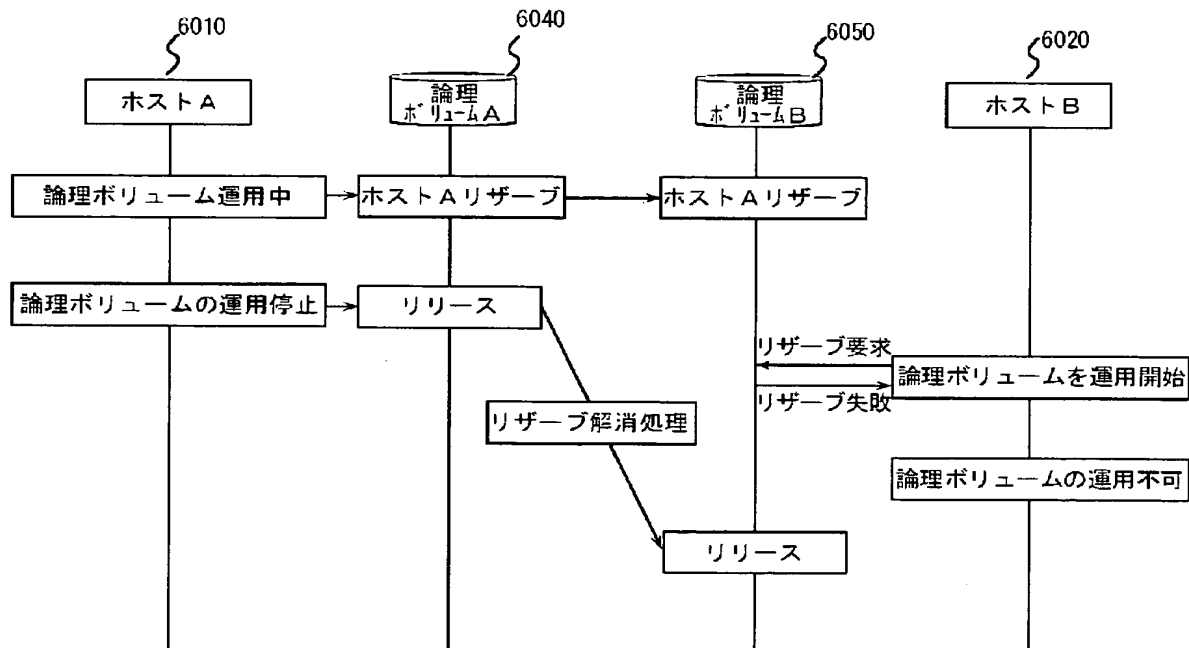
【図 22】

図 22



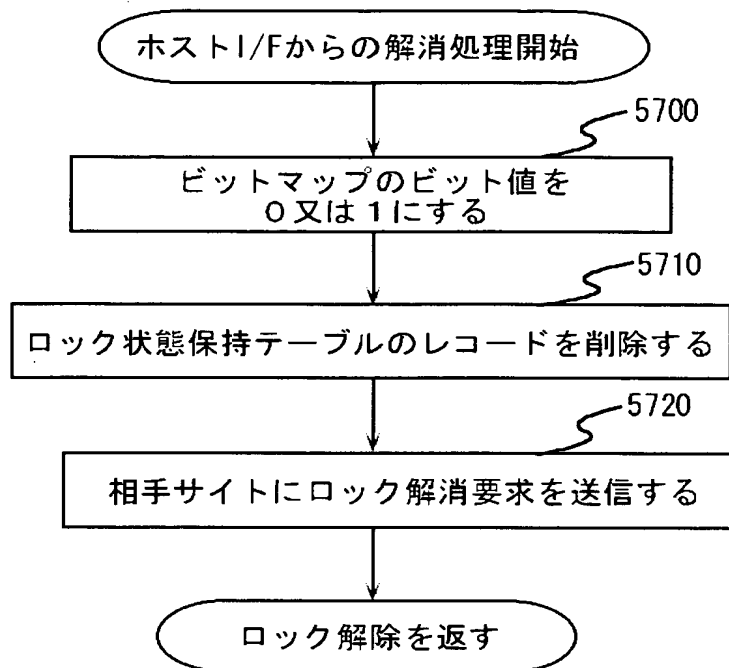
【図 23】

図 23



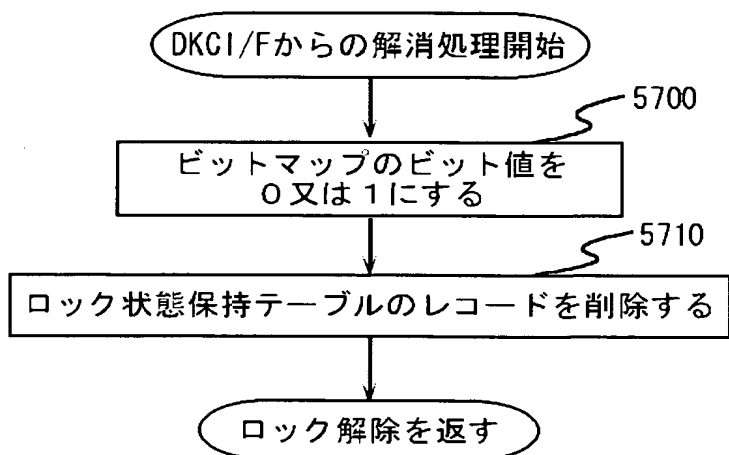
【図 2 4】

図 2 4



【図 2 5】

図 2 5



【書類名】 要約書

【要約】

【課題】 コピーペアを構成する記憶装置システム間でコピー方向を一方向に固定せずに双方向にコピーを行えるようにする。

【解決手段】 コピーペアを構成する記憶装置システム1070、1080は、データ一貫性保持制御部2040を設ける。データ一貫性保持制御部2040は、上位装置から受領した書き込みデータとDKCI/Fを介して他の記憶装置システムから受信した書き込みデータとが物理記憶装置の同一格納場所に重複して書き込まれるときに上位装置から書き込みデータを受領した受領時刻の順に書き込まれるように、コピーペアを形成する論理ボリュームへの書き込みデータを対応する受領時刻から所定時間以上キャッシュ部2050に待機させた後に物理記憶装置に書き込むよう制御する。

【選択図】 図4

特願 2 0 0 3 - 1 2 8 1 6 3

出 願 人 履 歷 情 報

識別番号

[0 0 0 0 0 5 1 0 8]

1. 変更年月日

1 9 9 0 年 8 月 3 1 日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台 4 丁目 6 番地

氏 名

株式会社日立製作所